

Automatic Analysis of People in Thermal Imagery

Liu, Jinsong

DOI (link to publication from Publisher):
[10.54337/aau499830013](https://doi.org/10.54337/aau499830013)

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Liu, J. (2022). *Automatic Analysis of People in Thermal Imagery*. Aalborg Universitetsforlag.
<https://doi.org/10.54337/aau499830013>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

AUTOMATIC ANALYSIS OF PEOPLE IN THERMAL IMAGERY

**BY
JINSONG LIU**

DISSERTATION SUBMITTED 2022



AALBORG UNIVERSITY
DENMARK

Automatic Analysis of People in Thermal Imagery

Ph.D. Dissertation
Jinsong Liu

Department of Architecture, Design and Media Technology
Aalborg University
Rendsburggade 14, 9000 Aalborg, Denmark
Dissertation submitted July, 2022

Dissertation submitted: July, 2022

PhD supervisor: Professor Thomas B. Moeslund
Aalborg University

PhD committee: Associate Professor Markus Löchtefeld (chairman)
Aalborg University, Denmark

Professor Alexandros Iosifidis
Aarhus University, Denmark

Professor Haibo Li
KTH Royal Institute of Technology, Sweden

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628

ISBN (online): 978-87-7573-872-4

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Jinsong Liu

Printed in Denmark by Stibo Complete, 2022

Curriculum Vitae

Jinsong Liu



Jinsong Liu was born on November 2, 1991 in Henan, China. He received the B.Eng. degree in Electronic Information Engineering from Henan University of Technology, Henan, China and the M.Eng. degree in Electronics and Communication Engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2012 and 2015, respectively. He is currently a PhD student at the Visual Analysis and Perception Laboratory, Department of Architecture, Design and Media Technology, Aalborg University, Aalborg, Denmark. His research interests include image processing and computer vision.

Curriculum Vitae

Abstract

As human beings, we are used to relying on ourselves to fulfill each task from our instinctive ability to see and understand the world by eyes and minds. However, the trend from manpower to automation in many industries has seen a great improvement in efficiency and performance, which encourages us to use the latest techniques to free ourselves from the labor of several works. This PhD work accordingly investigates the application of computer vision techniques to two projects *Thermal Adaptive Architecture* and *Safe Harbor* for the same expectation—automatic analysis of people in thermal imagery.

Thermal Adaptive Architecture is an indoor study that aims at building an office microclimate where individual occupant feels thermally comfortable, in which each person's clothing insulation rate I_{cl} and metabolic rate M have to be estimated. Therefore, we implement a tracking-by-detection module to track each individual, on top of which the key body parts are detected for measuring the skin temperature and clothes temperature that helps to calculate his or her I_{cl} . Besides, the detected bounding box of the person together with the optical flow in the box can describe the personal activity intensity from which M is estimated. Furthermore, inspired by the gender difference in thermal comfort assessment, we have done gender classification.

Safe Harbor is an outdoor study that aims at monitoring a harbor front and detecting anomalies of potentially dangerous or harmful incidents so that professionals can provide immediate controls or rescues. To this end, we propose an early-alarm strategy by detecting human activities in an alarm region that is very near to the waterside, to reduce the rescue preparation time for drowning accident prevention. Besides, we develop anomaly detection algorithms on a long-term thermal drift dataset that has very similar properties to a running surveillance system in real life rather than other short-term datasets that are far away from the real conditions.

The evaluation of the solutions on both projects has proven their feasibility in understanding humans automatically in thermal imagery. The complementarity of the indoor research and the outdoor research comprehensively explores the potential of using computer vision techniques to ease manual work for a more comfortable and safer life.

Abstract

Resumé

Som mennesker er vi vant til selv til at udføre opgaver ud fra vores instinktive evne til at se og forstå verden gennem øjne og sind. Tendensen fra manuel arbejdskraft til automatisering i mange industrier har set en stor forbedring i effektivitet og ydeevne, hvilket tilskynder os at bruge de nyeste teknikker til at frigøre os fra manuelt arbejde i endnu flere værker. Dette PhD arbejde undersøger derfor anvendelsen af computer vision teknikker indenfor to projekter; *Thermal Adaptive Architecture* og *Safe Harbor*, med samme forventning—automatisk analyse af mennesker i termiske billeder.

Thermal Adaptive Architecture er en indendørs undersøgelse der har til formål at understøtte et kontor mikroklima, hvor den enkelte beboer føler sig termisk komfortabel, hvor hver persons tøjisoleringsrate I_{cl} og stofskifte M skal estimeres. Derfor implementerer vi et tracking-by-detection modul til at spore hver enkelt person, hvorpå de vigtigste kropsdele detekteres til måling af hud temperatur og tøj temperatur, således at hans eller hendes I_{cl} kan beregnes. Desuden benyttes den detekterede bokse omkring personen sammen med det optiske flow i boksen til at beskrive den personlige aktivitets intensitet, ud fra hvilken M estimeres. Derudover har vi, inspireret af kønsforskellen i termisk komfort vurdering, lavet kønsklassificering.

Safe Harbor er en udendørs undersøgelse, der har til formål at overvåge en havnefront og opdage uregelmæssigheder i form af potentielt farlige eller skadelige hændelser, således at øjeblikkelig redning kan igangsættes. Til dette formål foreslår vi en tidlig alarmstrategi hvor menneskelige aktivitet detekteres, indenfor et alarmområde der er meget tæt på vandet, for at reducere forberedelsestiden for redning så drukneulykker undgås. Desuden udvikler vi anomalitets detektions algoritmer på et langsigtet termisk datasæt, der har meget af de samme egenskaber som et virkelighedens overvågningssystem, til forskel fra mindre tidsbegrænsede datasæts, der er langt væk fra de virkelige forhold.

Evalueringen af løsningerne på begge projekter har bevist deres gennemførlighed i at opnå en forståelse af mennesker automatisk igennem termiske billeder. Komplementariteten af den indendørs forskning og den udendørs forskning udgør en gennemgribende undersøgelse af potentialet ved at bruge

Resumé

computer vision teknikker til at lette manuelt arbejde for et mere behageligt og mere sikkert liv.

Contents

Curriculum Vitae	iii
Abstract	v
Resumé	vii
Thesis Details	xv
Preface	xvii
I Overview	1
1 Introduction	3
1 Automatic Analysis	3
2 Thermal Imagery	4
3 Research Scope	6
4 Reader's Guide	8
References	9
2 Thermal Adaptive Architecture	13
1 Background	13
2 Introduction	14
3 Related Work	17
3.1 Thermal Comfort Field	18
3.2 Computer Vision Field	20
4 Contributions	22
4.1 Data Collection	23
4.2 Personal Factors Estimation	28
References	42

3	Safe Harbor	49
1	Background	49
2	Introduction	50
3	Related Work	52
3.1	Drowning Accident Prevention	52
3.2	Anomaly Detection	54
4	Contributions	55
4.1	Data Collection	56
4.2	Drowning Accident Prevention: Warning in Advance . .	65
4.3	Anomaly Detection: on Long-term Data with Concept Drift	69
	References	76
4	Summary	81
II	Thermal Adaptive Architecture	85
A	Vision-based Individual Factors Acquisition for Thermal Comfort Assessment in a Built Environment	87
1	Introduction	89
2	Related Work	90
3	Proposed Method	91
4	Experiments	93
4.1	Dataset Information	93
4.2	Evaluation of the Proposed Method	95
5	Conclusions	98
	References	99
B	Automatic Estimation of Clothing Insulation Rate and Metabolic Rate for Dynamic Thermal Comfort Assessment	101
1	Introduction	103
2	Related Work	105
3	Proposed Method	108
3.1	Clothes and Activity Recognition	109
3.2	Skin and Clothes Temperatures Acquisition	110
3.3	I_{cl} and M Estimation	112
4	Experiments	114
4.1	Dataset Information	114
4.2	Evaluation of the CNN for Clothes Type and Activity Recognition	115
4.3	Evaluation of Temperature Acquisition	116
4.4	Evaluation of I_{cl} and M Estimation	119

5	Conclusions and Future Work	123
6	Appendix	124
	References	126
C	Clothing Insulation Rate and Metabolic Rate Estimation for Individual Thermal Comfort Assessment in Real Life	133
1	Introduction	135
2	Related Work	137
2.1	I_{cl} and M Estimation	137
2.2	Detection and Tracking	138
3	Methodology	139
3.1	Tracking-by-Detection	139
3.2	I_{cl} Estimation	143
3.3	M Estimation	147
4	Experiments	151
4.1	Dataset Information	152
4.2	Evaluation of the Tracking-by-Detection Module	152
4.3	Evaluation of the I_{cl} Estimation Module	155
4.4	Evaluation of the M Estimation Module	156
4.5	Application in Thermal Comfort Assessment	159
5	Conclusions and Future Work	159
	References	159
D	Knowing Where to Look: Gender Classification in Thermal Imagery	167
1	Introduction	169
2	Explainable CNN	171
2.1	Class Activation Mapping	171
2.2	Generating Class Activation Maps	173
3	Explainable Gender Classification	174
3.1	Dataset and Implementation Information	174
3.2	Results and Analyses	177
4	Discussion	181
5	Conclusion and Future Work	182
	References	183
III	Safe Harbor	189
E	Supervised versus Self-supervised Assistant for Surveillance of Harbor Fronts	191
1	Introduction	193
2	Related Work	195

Contents

2.1	Object Detection	195
2.2	Anomaly Detection	195
3	Challenges	196
3.1	Sensitive Data	196
3.2	Thermal Imaging	196
3.3	Rare Phenomena	197
4	Applied Methods	197
4.1	Supervised Human Detection	197
4.2	Self-supervised Anomaly Detection	199
5	Experiments	202
5.1	Supervised Surveillance Assistant	203
5.2	Self-supervised Surveillance Assistant	204
6	Discussion	207
7	Conclusion	207
	References	207
F	Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift	211
1	Introduction	213
2	Related Work	215
2.1	Concept Drift Detection	215
2.2	Datasets	215
3	The Long-term Thermal Drift (LTD) Dataset	216
3.1	Metadata Analysis	219
4	Long-term Performance Experiment	220
4.1	Data Selection Protocol	220
4.2	Tested Models	220
4.3	Drift Algorithmic Performance Analysis	221
5	Drift Analysis	223
6	Drift Prediction Baseline	224
7	Conclusion and Future Work	227
8	Appendix	228
8.1	Metadata Correlation	228
8.2	Data Sampling	229
8.3	Activity Calculation	230
	References	231
G	Detecting Anomalies Reliably in Long-term Surveillance Systems	241
1	Introduction	243
2	Related Work	245
3	Methods	246
3.1	Autoencoder	247
3.2	Background Estimation	247

Contents

3.3	Weighted Reconstruction Error	250
4	Experiments	251
4.1	Dataset Information	251
4.2	Implementation Details	251
4.3	Weighted MSE	252
5	Conclusions	258
	References	258
H Imitating Emergencies: Generating Thermal Surveillance Fall Data Using Low-Cost Human-like Dolls 263		
1	Introduction	265
2	Related Work	268
2.1	Fall Detection	268
2.2	Mannequins in Data Capture	269
2.3	Thermal Mannequins	270
3	Capturing Cameras	270
4	Thermal Doll Design	271
5	Doll Thermal Appearance	272
5.1	Comparing Temperature between the Doll and Real People	273
5.2	Doll Temperature Change over Time	275
6	Fall Motion Comparison	276
7	Doll Detection Comparison	279
8	Conclusions and Future Work	281
	References	282

Contents

Thesis Details

Thesis Title: Automatic Analysis of People in Thermal Imagery
PhD Student: Jinsong Liu
Supervisor: Professor Thomas B. Moeslund, Aalborg University

The main body of this thesis consists of the following papers and one technical report:

Thermal Adaptive Architecture

- [A] **Jinsong Liu**, Isak Worre Foged, and Thomas B. Moeslund. *Vision-based Individual Factors Acquisition for Thermal Comfort Assessment in a Built Environment*. The 1st Workshop on Faces and Gestures in E-health and Welfare (FaGEW) at the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020.
- [B] **Jinsong Liu**, Isak Worre Foged, and Thomas B. Moeslund. *Automatic Estimation of Clothing Insulation Rate and Metabolic Rate for Dynamic Thermal Comfort Assessment*. Pattern Analysis and Applications, 2021. <https://doi.org/10.1007/s10044-021-00961-5>
- [C] **Jinsong Liu**, Isak Worre Foged, and Thomas B. Moeslund. *Clothing Insulation Rate and Metabolic Rate Estimation for Individual Thermal Comfort Assessment in Real Life*. Sensors, volume 22, number 2, page 619, 2022. <https://doi.org/10.3390/s22020619>
- [D] **Jinsong Liu** and Thomas B. Moeslund. *Knowing Where to Look: Gender Classification in Thermal Imagery*. Technical Report, 2022.

Safe Harbor

- [E] **Jinsong Liu**, Mark P. Philipsen, and Thomas B. Moeslund. *Supervised versus Self-supervised Assistant for Surveillance of Harbor Fronts*. The 16th

International Conference on Computer Vision Theory and Applications (VISAPP) as part of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 2021.

- [F] Ivan Nikolov, Mark P. Philipsen, **Jinsong Liu**, Jacob V. Dueholm, Anders S. Johansen, Kamal Nasrollahi, and Thomas B. Moeslund. *Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift*. Datasets and Benchmarks Track at the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021.
- [G] **Jinsong Liu**, Ivan Nikolov, Mark P. Philipsen, and Thomas B. Moeslund. *Detecting Anomalies Reliably in Long-term Surveillance Systems*. The 17th International Conference on Computer Vision Theory and Applications (VISAPP) as part of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 2022.
- [H] Ivan Nikolov, **Jinsong Liu**, and Thomas B. Moeslund. *Imitating Emergencies: Generating Thermal Surveillance Fall Data Using Low-Cost Human-like Dolls*. *Sensors*, volume 22, number 3, page 825, 2022. <https://doi.org/10.3390/s22030825>

This PhD thesis is an article-based thesis that consists of published scientific papers and a technical report of ongoing work. The thesis is submitted for assessment in partial fulfilment of the PhD degree. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty.

Preface

This thesis is submitted to the Technical Faculty of IT and Design at Aalborg University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy. The thesis is constituted of two parts: an overview of the involved PhD projects *Thermal Adaptive Architecture* and *Safe Harbor* and a collection of published scientific papers (and a technical report). The overview is in Part I. The published papers (and the technical report) are in Part II and Part III corresponding to the two projects, respectively.

The project *Thermal Adaptive Architecture* was funded by Realdania Foundation and The Obel Family Foundation. The project *Safe Harbor* was funded by TrygFonden. Both projects gave me the opportunity to carry out the work from the end of 2018 to the beginning of 2022 in the Visual Analysis and Perception Laboratory (VAP) at the Department of Architecture, Design and Media Technology at Aalborg University.

Though the PhD work has come to its end, I have to admit that I could not have completed the work without my outstanding supervisor, smart collaborators, encouraging friends, and my beloved family. I thus first want to express my deepest gratitude to my supervisor, Prof. Thomas B. Moeslund, for his guidance, encouragement, trust, support, and care during my PhD study. He has given me the freedom to do what I am interested in and taught me that the nature of research is trial and error and never give up. All of these make me not only a good researcher but also a better man than before. Then I would like to thank Isak Worre Foged, my close collaborator and mentor on *Thermal Adaptive Architecture*, who provided me the chance to explore the world of architecture design which I had no clue about before commencing the project. This interdisciplinary research has helped to build my confidence that my competences in the field of computer vision can have an impact on many other fields as long as I have the desire to learn new knowledge. Further, I am thankful for all my colleagues at VAP, in particular, Mark P. Philipsen and Ivan Nikolov, who are my close collaborators in *Safe Harbor*, giving me great help to advance the progress of the project. As a foreign member at VAP, I have always been feeling welcomed and cared from my dear colleagues, especially during the dark COVID-19 days. Also, I

Preface

want to thank my supervisor during my master's study Prof. Shaosheng Dai and my previous colleague Cheng-Bin Jin for their help to guide me into the world of image processing and computer vision.

I would also like to express my thanks to my friends for their years of accompanying me, in particular, Xinchao Sun and Qiongxiu Li, who have made bad days good days, and good days great days for me. Though we are not blood relatives, in my deepest heart you two are already part of my family. Finally, I am especially thankful to my grandparents and my parents for their unconditional love for more than thirty years. Because of the confidence that they are always standing behind me and supporting me, I can face bravely all the ups and downs in my life. Now I am a grown-up man and it is my turn to shelter them from the wind and rain and bring them sunshine and happiness.

Jinsong Liu
Aalborg University, July 3, 2022

Part I

Overview

Chapter 1

Introduction

1 Automatic Analysis

Throughout the world history, from the Stone Age till the Industrial Revolution, humans had been relying on themselves to fulfill each task for earning a living, no matter making the simple stone tools or building the Wonders of the Ancient World that even astonish the modern civilization. From the 18th century, with the widespread applications of steam engine power, electricity, and petroleum energy, human manufacturing gradually turned to mechanization, leading to an unparalleled rise in productivity and the population growth rate. The shift from mechanical reliance to electronic and information technology which emerged from the middle of the 20th century has boosted the development in automatic computation, storage, and communication technology, making the internet, computers, and other smart devices part of daily life. The human history is running forward towards the Information Age. This reveals that the human society develops with a clear direction—from absolute manpower to full automation, to make life more convenient and make us feel more satisfied.

The ability to improve the world for humans to better live in cannot be separated from the understanding of the environment around us. As a kind of creature, we use our eyes to see what the world looks like and our minds to comprehend what it means to us. In an automated system, these two concepts correspond to cameras that capture scenes and computer algorithms that analyze the scenes. This camera-algorithm paradigm is the so-called computer vision that gains high-level understanding from digital images or videos, thus automating the tasks that are usually performed by the human visual system [1–3].

Therefore, nowadays, with computer vision algorithms, numerous tasks especially some repetitive and tedious assignments are going forward to au-

tomated ways at different levels, such as defect detection for clothes [4–6], grain diseases detection and identification in the food industry [7–9], biometric features recognition in access control in smart buildings [10–12], objects detection and road segmentation in automated driving systems [13–15], and crowds detection to control social distance in COVID-19 days [16–18].

Specifically, this PhD work applies computer vision to automatically analyse people in thermal imagery. Concretely, we have focused on two projects: *Thermal Adaptive Architecture* and *Safe Harbor*. *Thermal Adaptive Architecture* aims at understanding the thermal status of each occupant based on which the indoor microclimate can accordingly change to achieve individual thermal comfort state. *Safe Harbor* aims at monitoring a harbor region and detecting anomalies like drowning accidents that need timely controls or rescues by professionals. Both projects rely on thermal cameras as imagery acquisition hardware and algorithms as software to realize human detection, human tracking, behavior recognition, anomaly detection, etc., all of which will be introduced next.

2 Thermal Imagery

Cameras make themselves indispensable to our everyday life, especially with the popularization of mobile phones that are equipped with at least two cameras—a front-facing one and a rear-facing one. This seems to lead to a situation where each person has his or her camera to “see” the world. Almost with no exception, the cameras people are familiar with are visible cameras that convert visible light into RGB images or videos as those seen in cell-phones, computers, and television programs. However, taking pictures with a visible camera is greatly influenced by the illumination situation. Take a daily experience as a simple example that nothing can be seen and recorded by a visible camera in a totally dark environment.

Thermal cameras are developed and used to improve this limitation, bringing thermal imagery—a process of converting infrared radiation into visible images that describe the spatial heat distribution of a scene [19]. The mentioned infrared radiation can be emitted by any object whose temperature is higher than absolute zero. This temperature-induced property makes infrared radiation also named as thermal radiation [20]. Thereby there is no need for thermal imagery to rely on external illumination sources anymore as long as the object being captured emits its own thermal radiation.

The independence of external radiation and the reliance on objects themselves determine the widespread application of thermal cameras, such as animals detection in the wild as warm-blooded animals most often have a different temperature distribution from the surroundings [21], building inspection from outside as windows and doors have a higher heat loss than

2. Thermal Imagery

other regions [22], forest fire detection as hot spots [23], and of course human activity surveillance especially when there is bad lighting due to difficult weather conditions or periods at night [24].

Besides the above advantages of thermal imagery, other dominant strengths in terms of understanding people with thermal cameras are the privacy-friendly property and the ability to demonstrate a temperature distribution without contact. A thermal image with persons in it can largely increase the difficulty in indicating a person's identity than its visible imagery counterpart, which can be seen in Fig. 1.1 where the facial information is protected in thermal images (b) and (d). This not only lets humans under surveillance feel safe and not invaded but also complies with the General Data Protection Regulation (GDPR) in European Union (EU) [25]. On the contrary, a visible image needs further post-processing like blurring or pixelating identity information to keep each person's anonymity to comply with GDPR, which consumes extra time and computation resources [26].

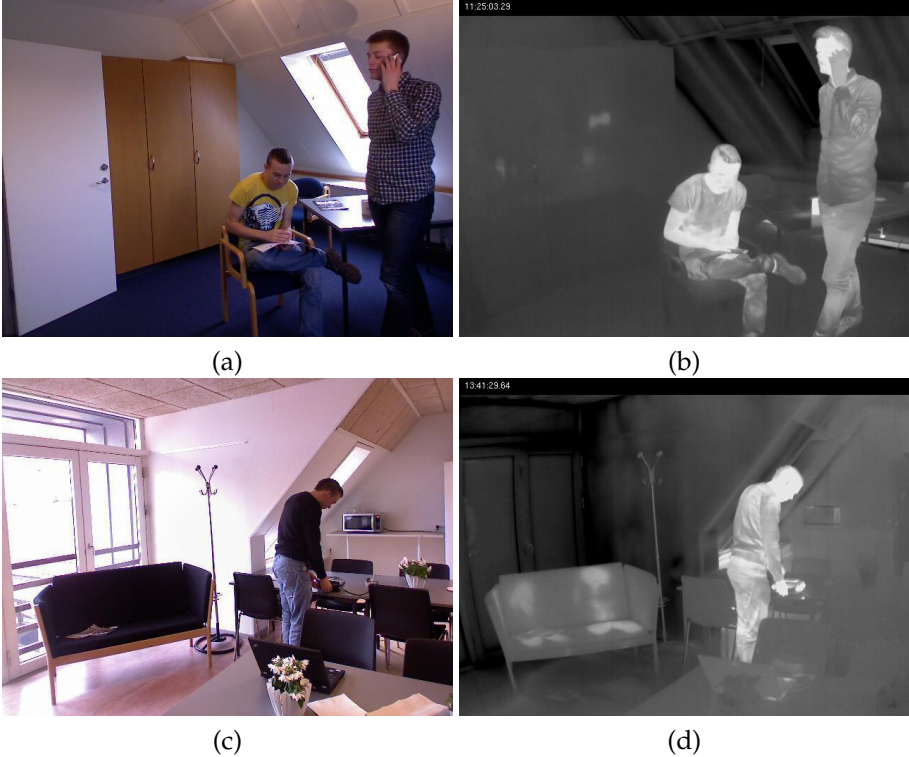


Fig. 1.1: Visible RGB images (a) and (c), and their thermal counterparts (b) and (d). Image source: [27].

When it comes to temperature visualization, since thermal imagery re-

sults from the thermal radiation generated from an object's heat distribution, regions with different temperatures will emit different intensities of radiation and then result in different grayscale values in the thermal image. As in Fig. 1.1 (b) and (d), brighter pixels indicate higher temperatures and darker pixels indicate lower temperatures; the notable temperature difference between a human and the background also eases the person region extraction for some tasks like human localization, especially in a complex environment full of chaotic objects like furniture in (a) and (c). This "what you see is what temperature you know" characteristic of thermal imagery is extremely popular in such COVID-19 days when many public places like airports employ thermal imaging cameras to detect passengers' forehead temperature to detect Coronavirus fevers, which further promotes the growth of thermal imagery market [28].

Therefore, considering all the strengths of thermal imagery, the two projects that the PhD work includes are using thermal cameras as hardware to capture scenes. In this way, the indoor research of *Thermal Adaptive Architecture* can protect each occupant's privacy information, and the outdoor research of *Safe Harbor* can resist the illumination shortage at night.

3 Research Scope

Although both projects, *Thermal Adaptive Architecture* and *Safe Harbor*, aim at understanding people in thermal imagery that determines some common research questions, the different research environments determine that the PhD thesis will have a relatively broad research scope to meet the individual requirements of the two projects. This complementarity of the indoor research and the outdoor research makes the PhD thesis more comprehensive to answer a general research question:

*To what extent can computer vision ease manpower
for a more comfortable and safer life?*

The word "comfortable" refers to an office microclimate that occupants are thermally satisfied with, corresponding to the following research questions in *Thermal Adaptive Architecture*:

- What factors influence a person's thermal sensation?
- How to acquire these factors from computer vision solutions?
- Compared with existing manual work, are the acquired factors from computer vision solutions sufficient and accurate?

3. Research Scope

The word “safer” refers to a harbor area where people feel safe to pass by and play in because anomalies—emergencies and potentially dangerous accidents—can be dealt with by professionals immediately, corresponding to the following questions in *Safe Harbor*:

- What anomalies should be considered and detected?
- How to detect the considered anomalies from computer vision solutions?
- Is the detection method fast and efficient for a timely control or rescue?

From the perspective of a computer vision researcher or engineer, to answer these questions, a series of general tasks need to be done, that is:

- **Data Acquisition** This refers to images or videos recorded from a sensor, i.e., a thermal camera in our case. For *Thermal Adaptive Architecture*, the thermal camera is recording the occupants’ daily office work. For *Safe Harbor*, the thermal camera is recording pedestrians, vehicles, animals, etc. who appear in the harbor area being monitored.
- **Feature Extraction** For further tasks of recognizing what an object is or localizing where it is, a prerequisite is to define what the object looks like. Similar to the situation where we humans use “round, red, and rising in the east and setting in the west” to describe the sun, a computer vision solution has to find and use a number of overall, local, spatial, and temporal features to accurately describe the object it is “observing”. The mentioned temporal features are those extracted across time instead of a single image frame, for example, the movement characteristic.
- **Object Classification** This refers to a task using the features that depict an object to predict what category the object belongs to, by comparing the similarity between the features and a set of predefined principles. For both projects, humans need to be discriminated from other objects via classification. For *Thermal Adaptive Architecture*, further categorizing a person’s clothes into long-sleeved or short-sleeved and categorizing a person’s body segments into the head, arm, hand, or other parts are required.
- **Object Detection** This refers to localizing all the objects of a certain category by a rectangular box around each object, which cannot be fulfilled without feature extraction, as explained before. For both projects, human detection is required before further automatic analysis.

- **Object Tracking** From the moment an object is detected the first time, a unique identity usually represented by a number is assigned to the object to track its movement in each frame of a video until it disappears. Therefore, with tracking, a certain object's location is always known in principle. For both projects, human tracking is greatly helpful for a deeper understanding of a particular person being observed.
- **Anomaly Detection** Anomaly detection refers to finding a rarely-happening event based on the recognized assumption that normality is what happens most frequently as the opposite of an anomaly. Therefore, directly comparing the similarity of an event to the principles of defining an anomaly or indirectly comparing the difference of it from the definition of a normal pattern can mark an anomaly out effectively. For *Safe Harbor*, rarely-occurring incidents having dangerous consequences are these we want to find via anomaly detection.

As a whole, the above research questions of the individual projects and the accompanying computer vision tasks will be studied in this PhD work. More precisely, there will be some differences between the two projects in solving the same task, and several tasks have been integrated into an end-to-end pipeline. The detailed information will be provided in the next chapters.

4 Reader's Guide

This PhD thesis includes three parts:

- Part I is an overview of the PhD work to guide the reader to grasp the key points from a convenient glimpse, which consists of four chapters: introduction, thermal adaptive architecture, safe harbor, and summary:
 - Chapter 1 gives a brief description of the topics and scopes the PhD thesis covers.
 - Chapter 2 describes the background, significance, targets, related work, and the contributions on *Thermal Adaptive Architecture*.
 - Chapter 3 describes the background, significance, targets, related work, and the contributions on *Safe Harbor*.
 - Chapter 4 summarizes the overview to emphasize the key findings of the PhD work and introduces the future work on both projects.
- Part II is a collection of three published scientific papers and one technical report of ongoing work on *Thermal Adaptive Architecture*.
- Part III is a collection of four published scientific papers on *Safe Harbor*.

The indoor research on *Thermal Adaptive Architecture* (Part I Chapter 2 and Part II) and the outdoor research on *Safe Harbor* (Part I Chapter 3 and Part III) are complementary to each other and together form a more comprehensive investigation of how computer vision technologies help realize an automatic analysis of people in thermal imagery.

References

- [1] Dana H Ballard and Christopher M Brown, *Computer Vision*, Prentice-Hall, 1982.
- [2] Thomas Huang, "Computer vision: Evolution and promise," 1996.
- [3] Milan Sonka, Vaclav Hlavac, and Roger Boyle, *Image processing, analysis, and machine vision*, Cengage Learning, 2014.
- [4] WK Wong and JL Jiang, "Computer vision techniques for detecting fabric defects," in *Applications of computer vision in fashion and textiles*, pp. 47–60. Elsevier, 2018.
- [5] Pandia Rajan Jeyaraj and Edward Rajan Samuel Nadar, "Computer vision for automatic detection and classification of fabric defect employing deep learning algorithm," *International Journal of Clothing Science and Technology*, 2019.
- [6] Aqsa Rasheed, Bushra Zafar, Amina Rasheed, Nouman Ali, Muhammad Sajid, Saadat Hanif Dar, Usman Habib, Tehmina Shehryar, and Muhammad Tariq Mahmood, "Fabric defect detection using computer vision techniques: a comprehensive review," *Mathematical Problems in Engineering*, vol. 2020, 2020.
- [7] Chia-Lin Chung, Kai-Jyun Huang, Szu-Yu Chen, Ming-Hsing Lai, Yu-Chia Chen, and Yan-Fu Kuo, "Detecting bakanae disease in rice seedlings by machine vision," *Computers and electronics in agriculture*, vol. 121, pp. 404–411, 2016.
- [8] Sourabh Shrivastava, Satish Kumar Singh, and Dhara Singh Hooda, "Soybean plant foliar disease detection using image retrieval approaches," *Multimedia Tools and Applications*, vol. 76, no. 24, pp. 26647–26674, 2017.
- [9] Tao Liu, Wen Chen, Wei Wu, Chengming Sun, Wenshan Guo, and Xinkai Zhu, "Detection of aphids in wheat fields using a computer vision technique," *Biosystems Engineering*, vol. 141, pp. 82–93, 2016.

References

- [10] Syafeeza Ahmad Radzi, MK Mohd Fitri Alif, Y Nursyifaa Athirah, AS Jaafar, AH Norihan, and MS Saleha, "Iot based facial recognition door access control home security system using raspberry pi," *International Journal of Power Electronics and Drive Systems*, vol. 11, no. 1, pp. 417, 2020.
- [11] Md Arif, Amdadul Haq, Md Zebon, Abdul Hye, and Sheikh Monzur Elahi, "Construction of a smart home automation system with voice activation and fingerprint security system," 2021.
- [12] R Rameswari, S Naveen Kumar, M Abishek Aananth, and C Deepak, "Automated access control system using face recognition," *Materials Today: Proceedings*, vol. 45, pp. 1251–1256, 2021.
- [13] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.
- [14] Ruturaj Kulkarni, Shruti Dhavalikar, and Sonal Bangar, "Traffic light detection and recognition for self driving cars using deep learning," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–4.
- [15] Yihuan Zhang, Jun Wang, Xiaonian Wang, and John M Dolan, "Road-segmentation-based curb detection method for self-driving via a 3d-lidar sensor," *IEEE transactions on intelligent transportation systems*, vol. 19, no. 12, pp. 3981–3991, 2018.
- [16] Fatma Bouhrel, Hazar Mliki, and Mohamed Hammami, "Crowd behavior analysis based on convolutional neural network: Social distancing control covid-19.," in *VISIGRAPP (5: VISAPP)*, 2021, pp. 273–280.
- [17] Onur Karaman, Adi Alhudhaif, and Kemal Polat, "Development of smart camera systems based on artificial intelligence network for social distance detection to fight against covid-19," *Applied Soft Computing*, vol. 110, pp. 107610, 2021.
- [18] Imran Ahmed, Misbah Ahmad, Joel JPC Rodrigues, Gwanggil Jeon, and Sadia Din, "A deep learning-based social distance monitoring framework for covid-19," *Sustainable Cities and Society*, vol. 65, pp. 102571, 2021.
- [19] Kirk J Havens and Edward Sharp, *Thermal imaging techniques to survey and monitor animals in the wild: a methodology*, Academic Press, 2015.

References

- [20] Rikke Gade and Thomas B Moeslund, "Thermal cameras and applications: a survey," *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.
- [21] Justyna Cilulko, Paweł Janiszewski, Marek Bogdaszewski, and Eliza Szczygielska, "Infrared thermal imaging in studies of wild animals," *European Journal of Wildlife Research*, vol. 59, no. 1, pp. 17–23, 2013.
- [22] JR Martinez-De Dios and Anibal Ollero, "Automatic detection of windows thermal heat losses in buildings using uavs," in *2006 world automation congress. IEEE*, 2006, pp. 1–6.
- [23] Isabelle-Gabriele Hendel and Gregory M Ross, "Efficacy of remote sensing in early forest fire detection: A thermal sensor comparison," *Canadian Journal of Remote Sensing*, vol. 46, no. 4, pp. 414–428, 2020.
- [24] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar, "Thermal object detection in difficult weather conditions using yolo," *IEEE access*, vol. 8, pp. 125459–125476, 2020.
- [25] European Parliament and Council of the European Union, "General data protection regulation," <https://gdpr-info.eu/>, 2016, last accessed: March, 2022.
- [26] Ivan Adriyanov Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Velling Dueholm, Anders Skaarup Johansen, Kamal Nasrollahi, and Thomas B Moeslund, "Seasons in drift: A long-term thermal imaging dataset for studying concept drift," in *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.
- [27] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmoose, Thomas B Moeslund, and Sergio Escalera, "Multi-modal rgb–depth–thermal human body segmentation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 217–239, 2016.
- [28] David Perpetuini, Chiara Filippini, Daniela Cardone, and Arcangelo Merla, "An overview of thermal infrared imaging-based screenings during pandemic emergencies," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, pp. 3286, 2021.

References

Chapter 2

Thermal Adaptive Architecture

1 Background

According to the World Health Organization, “health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” [1], emphasizing the health in mentality—inner peace, satisfaction, and happiness. To maintain such a completely healthy status, the improvement of workplace satisfaction for employees attracts more attention than ever before. Because in the current information era, people spend more hours indoors to fulfill tasks that are unprecedentedly dependent on computers and other electronic instruments. Though the satisfaction in a workplace should consider multiple facets like colleagues, facilities, personal security, etc., the workplace environment in terms of its indoor microclimate deserves special attention. Because the microclimate is the one that continuously influences an occupant from the moment he or she enters the working space until he or she gets off work.

In the project of *Thermal Adaptive Architecture*, the satisfaction with an indoor microclimate specifically refers to a thermal state where each person feels thermally comfortable in the office he or she works in. To achieve this thermal comfort status, adjusting the indoor microclimate adaptively according to the thermal conditions of occupants is an effective way. This adjustment is usually realized by controlling Heating, Ventilation, and Air Conditioning (HVAC) systems, the incident sunshine energy from glass windows like Fig. 2.1 shows, and other ways. Our daily experience that a remote controller can increase or decrease the chamber temperature easily as we want is such an example. Therefore, identifying and evaluating the real-time ther-

mal status of an occupant so that an architecture (an office, to be specific) can adaptively change its micro-environment is the idea underlined in this PhD project.

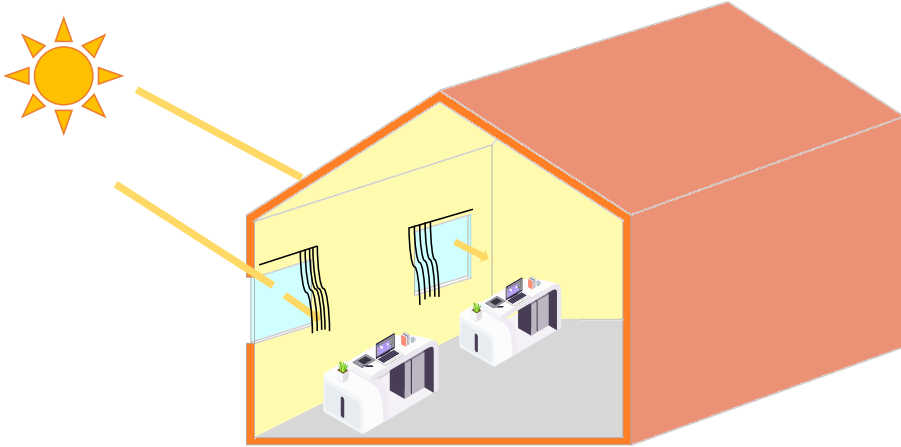


Fig. 2.1: Two curtains can separately control the incident sunshine energy for two office regions and thus provide different thermal conditions for two persons. If the curtain can be drawn automatically according to each person's thermal need, individual thermal comfort will be realized to some extent. The two office desks are icons downloaded from Iconfont © ALIMAMA MUX.

The necessity of research on *Thermal Adaptive Architecture* is not confined to a healthier mentality of a worker. On the one hand, thermal satisfaction with the working place can increase a person's working efficiency and productivity. On the other hand, HVAC-caused energy consumption in office buildings comprises a large proportion of the total energy used [2]. Therefore, abstemious energy consumption by the dynamic control of HVAC systems in offices according to real thermal needs can reduce the energy waste from overcooling or overheating, thus further reducing greenhouse gas emissions and contributing to the realization of the carbon neutrality goal.

2 Introduction

Again, the PhD project *Thermal Adaptive Architecture* is with the purpose of realizing the adaptation of an architecture microclimate to human thermal requirements so that each occupant in it (referring in particular to each individual in a certain office with an artificial climate) is thermally comfortable. For this purpose, a person's thermal sensation should be first known. Although this sensation can be expressed as simple as only three scales of hot, neutral, and cold that people often use on normal days, a standardized ASHRAE 7-scale thermal sensation representation is more recognized as listed in Table

2. Introduction

2.1 [3] where a scale of +1 (slightly warm), 0 (neutral), or -1 (slightly cool) is considered as an acceptable feeling of the environment generally.

Table 2.1: Seven-scale thermal sensations. The table is from [3] and also used in [4].

Sensation	Scale
Hot	+3
Warm	+2
Slightly warm	+1
Neutral	0
Slightly cool	-1
Cool	-2
Cold	-3

To quantify a person's thermal sensation to one of the seven scales, which is in another word called thermal comfort assessment, the first idea that comes to mind is using a questionnaire to record the feeling manually. However, this contradicts the principle of automatic analysis to ease manpower. Fortunately, the most important contributor in the thermal comfort field, professor Povl Ole Fanger, has proposed that the human thermal response to a certain environment is determined by four environmental factors and two personal factors [3, 5] as listed in Table 2.2.

Table 2.2: Six factors that determine a person's thermal sensation. Adapted from [3, 5, 6].

Environmental factors	Personal factors
Air temperature (T_a)	Clothing insulation rate (I_{cl})
Mean radiant temperature (\bar{t}_r)	Metabolic rate (M)
Air velocity (V_a)	
Relative humidity (RH)	

The four factors of T_a , \bar{t}_r , V_a and RH are physical properties describing an indoor microclimate. For clothing insulation rate I_{cl} , everyone has the experience that the clothes he or she wears on cold days are much thicker than those on warm days, which intuitively explains I_{cl} as an index to quantify the clothes' ability to insulate the bare skin from the outer air. Similarly, the daily experience that an activity with a higher intensity like running heats our body more effectively and quickly than taking a walk with a low intensity. This experience explicitly explains metabolic rate M as a numerical index of converting body chemicals like fat into thermal energy (that heats ourselves) and mechanical energy (that corresponds to the activity we are doing).

Generally, a thermal sensation is a combined effect of energy generation

of metabolism (related to M), energy loss by working on the surroundings via activity (related to M), and energy exchange between the person himself and the environment (related to T_a , \bar{t}_r , V_a , RH) via skin/clothes (related to I_{cl}) and breathing (related to M). Therefore, the assessment of a person's thermal sensation is possible as long as the six factors are acquired, by using them as inputs of Fanger's Predicted Mean Vote (PMV) model [7, 8]. Fig. 2.2 gives such an example of predicting the feeling scale as -0.4 that indicates a thermal neutrality state. In the figure, the red circle represents the exact combination of the provided six parameters listed in the left column, and the blue region represents all the combinations that can produce a comfortable feeling corresponding to a PMV from -0.5 to +0.5. This narrow range of $[-0.5, +0.5]$ is even more strict than the normally accepted range of $[-1, +1]$ mentioned in the description of Table 2.1, for the reason that the PMV was initially designed as the index for predicting the mean thermal response of a group of people exposed to the same indoor environment. In this context, a PMV value from -0.5 to +0.5 can fulfill the criteria of a thermally comfortable chamber—at least 90% of the persons in it are thermally satisfied, or in other words, a Predicted Percentage of Dissatisfied (PPD) rate is lower than 10%.

But the initial design for a group of people does not exclude the application of the PMV model to individual thermal comfort assessment, which is especially important in *Thermal Adaptive Architecture* where “individual” is the most distinctive highlight from other similar studies. That means the indoor environment in *Thermal Adaptive Architecture* is designed to be suitable for every individual by separately controlling local thermal conditions. For this goal, the four environmental factors of each local space and the two personal factors of each person have to be measured. Among them, multiple thermometers, anemometers, and hygrometers installed in separate locations according to ISO standard 7726 [11] can provide T_a , \bar{t}_r , V_a and RH values easily. Nevertheless, the measurement of I_{cl} and M is usually done manually by comparing an individual's clothes type and activity type with lookup tables [3, 8, 12–18] (partly shown in Table 2.3 and Table 2.4) to get the reference values, which is time-consuming, inefficient, and far from convenient. This points out the necessity for developing a faster and more convenient method to estimate I_{cl} and M , especially considering that the two personal factors are with a dynamic property and thus constantly change.

In summary, to realize the adaptive adjustment of the architecture's microclimate according to each person's thermal need in *Thermal Adaptive Architecture*, from the perspective of the PhD work, the key research question is how to dynamically acquire a person's clothing insulation rate I_{cl} and metabolic rate M with a thermal camera as hardware and computer vision algorithms as software.

3. Related Work

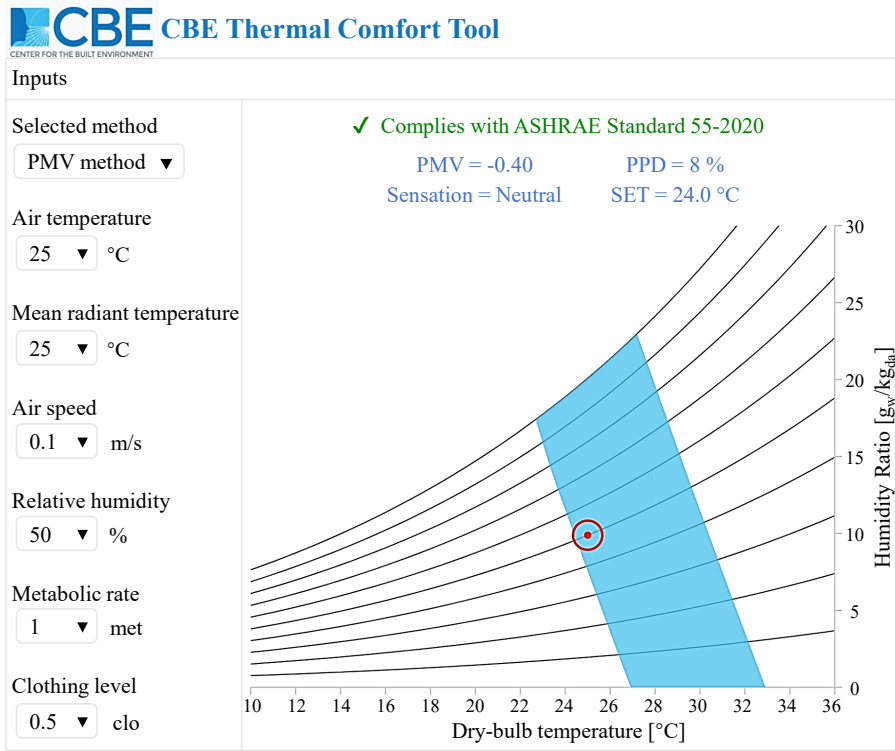


Fig. 2.2: An example of thermal comfort assessment with the six factors as inputs by using PMV model realized in CBE thermal comfort tool [9]. The figure is adapted from [10] for better visualization.

Table 2.3: Clothing insulation rates of various typical garments [13]. The table is also used in [6].

	Garment	I_{cl} (clo)
Underwear	Singlet	0.04
	T-shirt	0.09
	Shirts with long sleeves	0.12
Shirts, blouses	Short sleeves	0.15
	Lightweight, long sleeves	0.2
	Normal, long sleeves	0.25

3 Related Work

Thermal Adaptive Architecture is an interdisciplinary project between the thermal comfort field and computer vision field. For a convenient guide for

Table 2.4: Metabolic rates ($58 \text{ W/m}^2 = 1 \text{ MET}$) of typical activities [14]. A similar table is used in [4].

Activity	$M \text{ (W/m}^2\text{)}$
Sleeping	40
Reclining	45
At rest, sitting	55
At rest, standing	70
Walking on the level, even path, solid:	
1. without load	
2 km/h	110
3 km/h	140
4 km/h	165
5 km/h	200
2. with load	
10 kg, 4 km/h	185
30 kg, 4 km/h	250

readers to the solutions to acquire an individual's clothing insulation rate I_{cl} and metabolic rate M , in the section, only representative studies from both domains are introduced, provided that the work can ease the manpower dedicated to the lookup tables mentioned above. Additional related works are introduced in the scientific papers appended in Part II.

3.1 Thermal Comfort Field

In this field, clothing insulation rate I_{cl} estimation and metabolic rate M estimation are always studied in separate publications.

Clothing Insulation Rate Estimation

When choosing the clothes to wear, people usually refer to the outdoor climate. This leads to an assumption that people need clothes with low I_{cl} (a typical value of 0.5 clo) on warm days and clothes with high I_{cl} (a typical value of 1.0 clo) on cold days, which is considered to be the simplest I_{cl} estimation scheme mentioned in [3]. However, we all have to admit that a general outdoor climate cannot represent the specific thermal condition of an indoor environment such as an office in our project. To this end, more factors have been taken into account. For example, [19] collects 1316 sets of different participants' personal information (I_{cl} , gender information, and transportation mode) and the corresponding non-personal information (outdoor air temperature at 6:00 a.m., dew point temperature at 6:00 a.m., and

3. Related Work

season) through questionnaire surveys to investigate the numerical relationship between the I_{cl} value and the other five parameters, which makes it possible to predict a person's I_{cl} if the other information is collected. Similarly, research [20] further considers the influence of indoor temperature, air conditioners, and latitude on people's dress behavior. These studies are much more flexible than the fixed I_{cl} values of 0.5 clo and 1.0 clo and have relieved the manpower in terms of looking up tables to some extent, but collecting new information through questionnaires may lead to extra manual work. What's more, in real life, individuals with exactly the same factors considered in [19] and [20] may have very different clothes choices due to various personal preferences, indicating that the I_{cl} estimation should not be decoupled from the clothes themselves.

A common sense that heavier clothes have a better protection ability against cold air gives an impression of the correlation between the clothes mass and the I_{cl} value, and hence encourages the work [21] to accordingly estimate I_{cl} by weighing the clothes. However, as pointed out by [13] that "garment mass on its own is not a good predictor of garment insulation", studies on this scheme are quite few. An equally widely accepted common sense that the temperature difference between the inner surface and the outer surface of an item seems to represent its ability to insulate heat. Such an example is that a thermos cup full of cold or hot water has a much larger inside-outside difference in temperature than a regular cup made of glass, and the content in the thermos cup can remain cooler or hotter than the outer surroundings for a much longer time period. This common phenomenon in daily life tallies well with the insulation rate calculation defined in ISO 7933 [12] and ISO 9920 [13], according to which the skin temperature and the clothing temperature are required to be measured. Work [22], therefore, uses two infrared thermopile sensors to measure the temperature of the neck or the ankle and the temperature of clothes to calculate I_{cl} .

Metabolic Rate Estimation

According to ISO 8996 [14], higher levels of M determinations are suggested by measuring the subject's heart rate, oxygen consumption and carbon dioxide production rate, or energy expenditure from the body to the environment. On the one hand, these methods work on the basis that the interior metabolic rate is often exposed as the explicit activity intensity. The higher the intensity, the higher the heart rate and the faster the breath to inhale more oxygen and exhale more carbon dioxide. On the other hand, in terms of the thermal effect, a higher internal metabolic rate transfers more heat to the environment, which can be measured if the chamber is confined.

Therefore, corresponding studies have emerged. In [23] and [24], a Fitbit Charge HRTM wearable wristband [25] and a BioHarness 3.0 fixed on the

chest with a strap [26] are used to monitor a user’s heartbeat for measuring M , respectively. A COSMED K5 device [27] is used in [28] and an AE-100i device from East Medic [29] is used in [30] for the same purpose—measuring a user’s oxygen consumption and carbon dioxide production to monitor the metabolic rate M . Such devices have much better accuracy in M estimation than the lookup table counterpart, mainly because they provide the feasibility of measuring a person’s dynamic metabolic rate in real time. However, the inconvenience that a user has to wear the apparatus on the wrist or the chest and even a mask (see Fig. 2.3) on the face heavily impedes this hardware-relied M determination in everyday life.

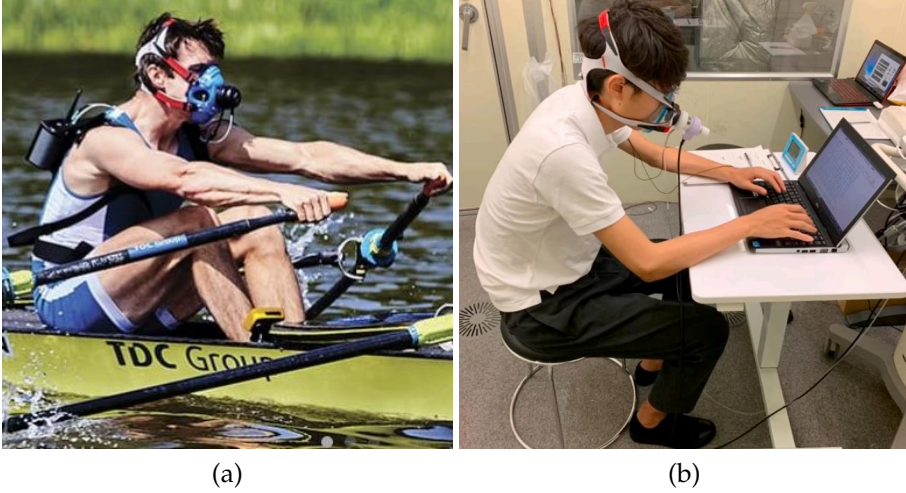


Fig. 2.3: Devices to measure oxygen consumption and carbon dioxide production. (a) is the device of COSMED K5 and (b) is the device of AE-100i. Image source: (a) [27], (b) [30].

3.2 Computer Vision Field

The above estimations of I_{cl} and M in the thermal comfort field are realized by questionnaires, attached sensors, or wearable devices, which makes it impossible for a subject to do things without distraction. Therefore, contactless approaches are much more welcomed—computer vision solutions benefited from the employment of cameras.

Clothing Insulation Rate Estimation

A thermal camera can directly output the temperature of each point in the scene it captures, therefore, this provides a contactless way to acquire a person’s skin temperature and clothes surface temperature which are the crucial

3. Related Work

values in calculating I_{cl} as explained in ISO 7933 [12] and ISO 9920 [13]. Accordingly, this convenient measurement in temperature has been adopted by studies [31–33] to estimate I_{cl} . As shown in Fig. 2.4, the polygon in (a) represents the clothes region; the circles in (b) represent the corresponding skin region and clothes regions; temperatures of these regions can be read directly from the thermal camera. However, an unsolved problem is that all the three publications do not mention how to locate skin regions and clothes regions automatically. That means if the polygon in Fig. 2.4(a) and the circles in Fig. 2.4(b) have to be drawn by humans, the introduced manual labor will contradict the expected automatic analysis.

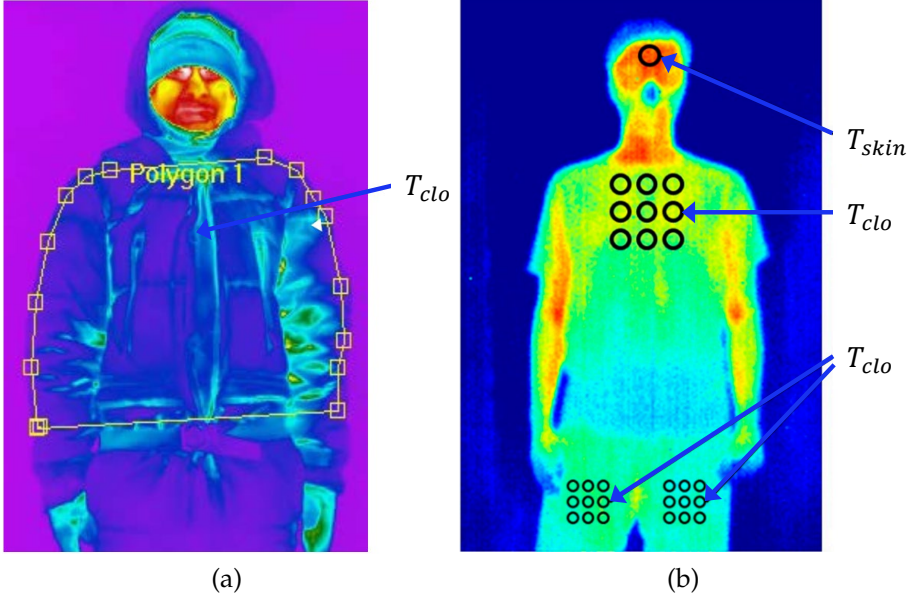


Fig. 2.4: Acquisition of skin temperature T_{skin} and clothing temperature T_{clo} in pseudo color thermal images captured by thermal cameras. Image source: (a) [32], (b) [33].

A different scheme to estimate I_{cl} by computer vision is from [34]. This work first chooses five types of clothing ensembles with known insulation rates (0.2 clo, 0.35 clo, 0.5 clo, 0.65 clo, and 0.8 clo) and takes 1000 pictures of such clothes. And then [34] uses these images to train a clothing ensemble classifier on the basis of a Convolutional Neural Network (CNN) so that the classifier gains the ability to categorize an image of clothes into one of the five types. In the testing/application phase, the image of a person wearing any clothes is taken by an RGB webcam, and then the image is fed to the classifier as the input to get its predicted category which corresponds to a known I_{cl} value. Apparently, the limitation is that only five types of clothing ensembles cannot represent so many clothes choices in our life. It is worth

mentioning that the brief description of this study—collecting a large number of images of known categories, training a CNN classifier by “observing” the images, using the well-trained classifier to predict the category of any testing image—is the typical pipeline of object classification in current computer vision solutions.

Metabolic Rate Estimation

Vision-based solutions to estimate M are rare. Studies [35, 36] use a Microsoft Kinect camera to capture the image information of a person doing various activities. At the same time, a Fitbit Charge wearable wristband the person wears records the heartbeat, letting each captured image correspond to a value of the heartbeat. And thus, numerous such pairs can infer the relationship between the image of a person doing a certain activity and the heart rate value of the person, with the help of a CNN-based regression model. As long as the regression relationship is determined, a subject does not need to wear the wristband anymore. Because with the image of him or her captured from Kinect as the input of the regression model, the heartbeat is predicted as the model’s output based on which the metabolic rate M is calculated following ISO 8996 [14].

4 Contributions

From what has been introduced, the goal of the PhD project *Thermal Adaptive Architecture* is to estimate each individual’s clothing insulation rate I_{cl} and metabolic rate M dynamically and automatically by computer vision solutions equipped with a single thermal camera. This goal can further facilitate the subsequent processing by our collaborators from the architecture design field to assess individual thermal sensations based on which to control the office microclimate so that each office worker feels thermal comfort.

From existing studies, we find that even though some of them have applied contactless computer vision algorithms in estimating I_{cl} or M , they are far away from the goal that *Thermal Adaptive Architecture* expects, especially as to the application in real life, mainly due to these inadequacies:

- The scheme from a certain clothes type to its predefined I_{cl} only considers a very limited number of garment types, which cannot represent hundreds of clothes choices in the real world.
- The scheme to calculate I_{cl} from a person’s skin temperature and clothing temperature does not give a solution to measure these temperatures automatically. That is to say, it is unable to locate the skin region and the clothing-covered region from an algorithm’s ability.

4. Contributions

- The scheme to determine M from a person’s heartbeat does not consider the individual difference in heart rates (some people are born with faster or slower heartbeats than others). And the need for an additional device to calibrate the paired relationship between an image and the value of heartbeat may involve some interference in subjects than a contactless camera.
- Most importantly, all existing approaches only take a single person as the research object, which is incapable of analyzing multiple persons captured in an image.

To this end, this PhD work develops new solutions to solve the above-mentioned drawbacks so that we can employ our method in a real-life office environment with a single privacy-preserving thermal camera. Below, the different components of our work during the PhD study are described.

4.1 Data Collection

Similar to the experience that a small child learns to recognize animals by repeatedly looking at pictures of various species of animals, current computer vision algorithms are also mainly built on the data-driven concept that needs a very large number of images or videos. Likewise, the knowledge about the animals that the child gets in the mind is like the parameters that a computer vision algorithm/model consists of. Therefore, to obtain the model that is able to recognize and locate a person, distinguish different types of clothes, and categorize various activities, two datasets have been collected during the PhD study.

Single-person Dataset [4, 37]

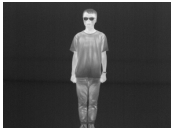
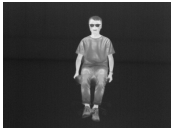
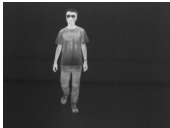
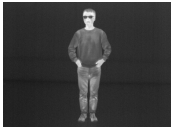





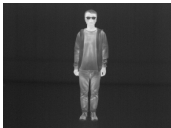


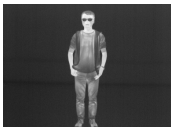


This dataset was collected in September 2019. Sixteen males and four females were asked to act three behaviors in five types of clothes as listed in Table 2.5, in front of a thermal camera—Xenics Gobi-384-GigE [38] demonstrated in Fig. 2.5. This collection consists of 291 long videos, and then they were trimmed into 2422 short videos. Each short video has a frame rate up to 30 fps, an image resolution of 384×288 pixels, and a length of around 3.5 seconds.

The principles for considering the five types of clothes and three types of behaviors are:

- An office worker’s lower body part is usually occluded by the desk in front, and thus a camera cannot capture the lower body information. Therefore, the clothes types here only consider the garment worn on the upper body.

- The five clothes types—T-shirt, with long sleeves, with rolled-up sleeves, with long sleeves and unzipped zipper, with rolled-up sleeves and unzipped zipper—can almost cover all the clothes situations in daily life.
- Sleeves conditions are good indicators of skin regions and clothing-covered regions. For example, the lower arms are usually considered as the bare-skin region when a person is wearing short-sleeved clothes.
- Zipper conditions are good indicators of clothes layers that may influence clothing insulation rate estimation according to ISO 9920 [13].
- Rolling up one's sleeves or not and unzipping one's zipper or not are direct information indicating a thermal feeling of hot or cold.
- Standing, sitting, and walking are the dominant behaviors in an office environment.

Table 2.5: Sampled images from the single-person dataset of people acting three different behaviors and in five different types of clothes.

Clothes	Behaviors		
	Stand	Sit	Walk
T-shirt			
Long sleeves			
Rolled-up sleeves			
Long sleeves & Unzipped zipper			
Rolled-up sleeves & Unzipped zipper			

4. Contributions



Fig. 2.5: The used thermal camera Xenics Gobi-384-GigE. Image source: © Xenics.

In each of the 2422 videos, there is only one person acting a certain behavior and wearing a certain type of clothes, as the illustration in Table 2.5. Therefore, each video has a specific label annotating the clothes property and behavior property of the person being recorded. Together there are 15 categories of videos considering the 15 combinations of five clothes types and three behavior types. This mix of both types makes it possible to recognize the two properties simultaneously instead of two separate pipelines, which is more convenient and efficient in practice.

In summary, our contributions regarding this single-person dataset are:

- We have collected a new dataset for analyzing a single person in thermal imagery. This dataset consists of 2422 videos in which each video has been labelled as a specific category that annotates the clothes and the behavior situation of the recorded person.
- Part of the dataset has been made public at <https://www.kaggle.com/datasets/jsliu91/single-person-thermal-dataset>, for other researchers and engineers to use. The remaining part of the dataset is for internal use to protect the face information of the participants who are reluctant to appear online.

Other information of this single-person dataset is in paper A [37] and paper B [4] appended in Part II.

Multiple-person Dataset [6]

Though the single-person dataset almost considers all the clothes and behavior situations in an office and thus fulfills the need of the early-stage research

during the PhD study, we claim that it cannot fully simulate the real working environment due to the following reasons: (1) the situation where an upper body is occluded by computer monitors or other stuff in front is not considered; (2) the situation where multiple persons are captured in the same video when they are working in the same office is not considered; (3) a person's clothes and activity situation may be changing in real time, but the video-level recognition (instead of a frame-by-frame recognition) cannot meet this requirement.

Therefore, in December 2020, we proposed new protocols on top of the important principles of the single-person dataset and accordingly collected a new multiple-person dataset, so that the real office working condition can be truly simulated. These new protocols are:

- Two persons were recorded together in a typical office environment with laptops or desktop monitors in front.
- Both persons were encouraged to behave spontaneously as they were doing regular work. This made a variety of behaviors like discussing with each other, taking notes, typing the keyboard, phoning, drinking water, stretching arms, etc. available in the new videos.

Under these protocols, three females and seven males were the subjects who helped us finally collect 114 videos by the same thermal camera as used for the single-person dataset. Each video has a frame rate up to 25 fps and a length of about 2000 frames. In this new multiple-person dataset, the two persons in each video can wear different types of clothes and act different activities. Besides, they may change their behaviors in real time. Therefore, each person needs to be analyzed separately in a frame-by-frame level.

For the frame-by-frame analysis of people, each person has to be detected, tracked, and recognized to get his or her clothes and activity situation. Accordingly, the multiple-person dataset has to be annotated to provide such information for training and evaluating a model. We thus sampled 5263 images that are evenly distributed from the 114 videos, and then we labelled each person's location with a bounding box around and the category to represent his or her clothes type and key posture in all the sampled images, making this dataset a good resource for a frame-by-frame analysis of any person in it.

For better illustration, Fig. 2.6 shows how the frames are annotated and what the videos look like. In the figure, people are demonstrating different kinds of activities like typing (a), texting (b), reading (c), drinking (d), stretching arms (e), and chatting (f). The green bounding box around each person labels the location information, and the accompanying category labels his or her clothes and behavior status.

It is worth mentioning that: (1) the clothes type is represented by the sleeves status as the single-person dataset does, but the zipper status is not

4. Contributions

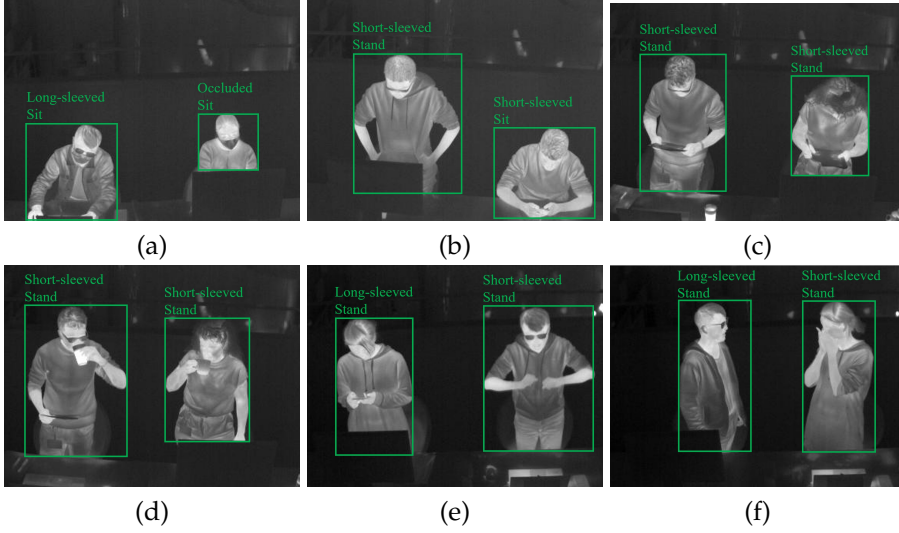


Fig. 2.6: Some images from the videos of the multiple-person dataset. During the data collection, participators were behaving in various activities without any restrictions. Also, each person in the sampled frames has been annotated with his or her location, clothes type, and key posture.

considered anymore for simplicity, so more attention is paid to the emphasis that this new dataset is for expanding the research to a real-life multi-person scenario; (2) a new clothes status called “occluded” is for the scene that the person is occluded by stuff in front, and thus it is challenging to know whether the garment is long-sleeved or short-sleeved; however, although a person may be occluded in some frames, it is possible to know the sleeves type from other frames; (3) there are two postures (sitting and standing) without walking considered in this multiple-person dataset, as walking is a particular type of standing and can be recognized in a frame-level analysis if a person’s location is constantly changing while standing.

In summary, our contributions regarding this multiple-person dataset are:

- We have collected a new dataset for analyzing multiple persons in thermal imagery. This dataset consists of 114 videos in which each video has two persons acting different activities in various clothes in an office. Besides, 5263 frames have been sampled from the videos and then annotated of each person’s location, clothes type, and posture, making the dataset a good resource for many tasks in the thermal mode.
- Part of the dataset has been made public at <https://www.kaggle.com/datasets/jsliu91/multiple-persons-thermal-dataset> so that other researchers and engineers can use. The remaining part of the dataset is for internal use to protect the face information of the participants

who are reluctant to appear online.

For more detailed information on this multiple-person dataset, please refer to paper C [6] in Part II.

4.2 Personal Factors Estimation

This section introduces our investigations and findings in estimating the two personal factors that affect individual thermal comfort assessment— I_{cl} and M , by recalling the key points in publications A [37], B [4], and C [6] listed in Part II. These works are specially characterized by that they are trying to estimate both factors simultaneously to increase convenience and lower the consumption in computation and processing time, which is very different from the existing works. Besides I_{cl} and M estimation, we have also studied an additional topic—gender classification in thermal imagery. This study is inspired by the gender difference in thermal comfort assessment. The work is ongoing and therefore documented in a technical report D appended in Part II. Below, the brief introduction of the four works will be described.

A: Vision-based Individual Factors Acquisition for Thermal Comfort Assessment in a Built Environment [37]

The first step of estimating I_{cl} and M is to predict an individual's clothes and activity situation, for which this work has been done to concurrently recognize a person's clothes type and behavior type on the collected single-person dataset.

Specifically, we have implemented a CNN to do a 15-category video classification task that will indicate what the person in the video is wearing and behaving based on the predicted class, according to the descriptions of Table 2.5. This CNN takes both spatial information (each thermal video) and temporal information (optical flows extracted from the thermal video) as the input for the reason that spatial information indicates the clothes type and temporal information indicates the behavior-related movements.

Different from prior works [39–41] that use two CNNs to separately deal with the video stream and the optical flow stream, our work incorporates the spatial-temporal information into one stream by concatenating each thermal image (8-bit) with its two optical flows (extracted from the horizontal direction and the vertical direction) into a 24-bit modality. This concatenation strategy not only saves resources and improves efficiency but also fills the gap that a temporal stream does not include clothes information.

Besides, to further increase the video classification accuracy, we take advantage of the long-term information of each video by doing result fusion inspired by [41]. This fusion is realized in three steps: (1) each thermal video

is segmented into K parts of the same length; (2) each segment randomly samples out one frame as the input of the CNN-based 15-category classifier; (3) the classification scores of the K sampled frames are averaged based on which the final class is predicted. In this way, the classification result considers the long-term information of each video with the help of the K frames that are sampled from the video along with time.

The diagram illustrating the above processing pipeline is in Fig. 2.7 where the concatenated spatial-temporal information and the result fusion strategy are drawn. This pipeline achieves an average recognition accuracy of 95.14% in a 15-category (the five clothes classes and three behavior classes listed in Table 2.5) classification problem.

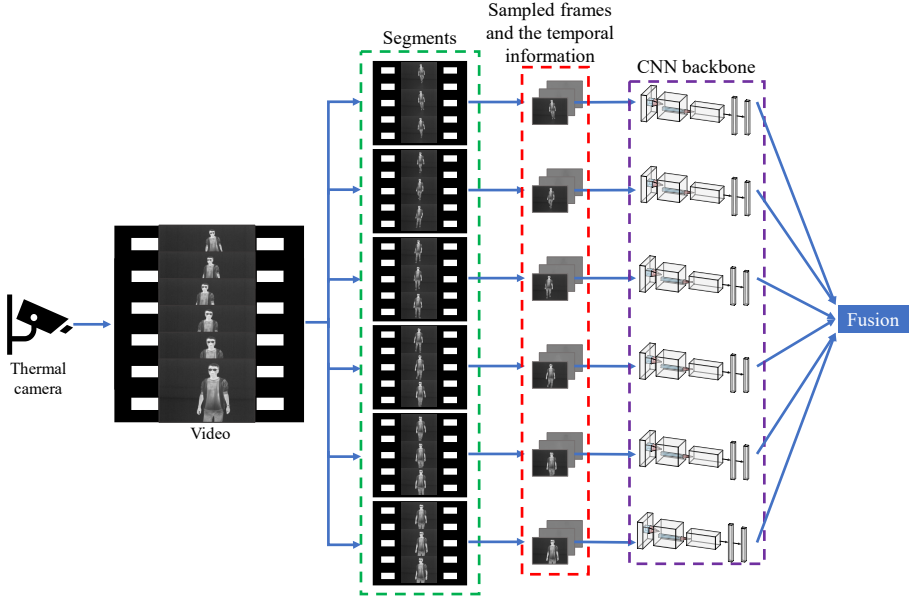


Fig. 2.7: The diagram of the work A. We use the optical flows of the horizontal and vertical directions to represent the temporal information. The result fusion strategy is to use the long-term information for better performance. To be noted is that though a few CNN backbones are drawn, there is only one CNN implemented in the whole pipeline. Therefore, this fusion strategy does not consume more resources. Adapted from [37].

In summary, our contributions regarding this work A are:

- We have implemented a CNN-based video classifier that considers both spatial information and temporal information to simultaneously recognize a person's clothes type and behavior type, which makes it possible to realize further I_{cl} and M estimation. As far as we can tell, this is the first work that acquires the individual factors for thermal comfort assessment from a perspective of clothing (sleeves and zipper) situations

and behaving actions.

- We have done extended experiments to discover some good practices for similar tasks involving optical flow extraction and thermal image recognition: (1) the camera should avoid a too high frame rate, because under this situation there is no visible movement between two adjacent frames, which makes the extracted optical flow information almost empty; (2) sometimes, a proper preprocessing step for a thermal input is necessary to enhance the objects due to the much fewer details compared with an RGB video.

For more detailed information of this work A, please refer to paper A in Part II.

B: Automatic Estimation of Clothing Insulation Rate and Metabolic Rate for Dynamic Thermal Comfort Assessment [4]

This work is built on top of paper A and aimed at estimating I_{cl} and M from the recognized clothes type and behavior type.

For I_{cl} estimation, the difference between an individual's skin temperature and clothes temperature has been proven crucial according to ISO standards [12, 13]. And thus, automatically locating the skin region and clothing-covered region is extremely important for extracting the corresponding temperatures, which includes two sub-problems: (1) how to distinguish the two regions from each other; (2) how to locate a certain point in the known skin region or clothes region. To solve the first sub-problem, from the recognized clothes type of short sleeves or long sleeves, we can treat the lower arm as the skin region or not, thus increasing the variety and robustness compared to the existing work which only considers facial area as the skin region. To solve the second sub-problem, we have applied OpenPose [42] to the single-person dataset to detect key body parts like shoulders, elbows, wrists, etc. These key parts are good landmarks to locate a skin-region point or a clothes-region point. For example, the shoulders are always covered by clothes, while the wrists of a person in short-sleeved clothes are the points in the skin region. By taking advantage of a thermal camera, the temperatures of these landmarks are easily acquired and can be good representations of the skin temperature and clothes temperature. Moreover, after calculating the I_{cl} value based on the above-acquired temperatures, we have further considered more factors that may have an influence on I_{cl} according to ISO 9920 [13], in order to make the estimated I_{cl} value more accurate. These factors include: (1) the decrease in clothing-covered body surface ratio caused by rolled-up sleeves or unzipped zippers will induce a decrease in I_{cl} ; (2) the extra insulation from an office chair when sitting will induce an increase in I_{cl} ; (3) the

4. Contributions

increased air exchange via clothes collars and cuffs when walking will induce a change in I_{cl} .

For M estimation, the recognized behavior of sitting, standing, or walking is linked to a specific metabolic rate value according to ISO 8996 [14]. This scheme provides an automatic lookup without manual work and any additional device. Considering that these three types of behaviors dominantly occur in an office environment, they can efficiently represent an office worker's M value as ISO 8996 [14] always does.

The pipeline of this automatic estimation of I_{cl} and M on the single-person dataset is illustrated in Fig. 2.8 where the module for recognizing clothes type and behavior type, the module for acquiring skin temperature and clothes temperature, and the module for calculating I_{cl} and M are shown. From the figure, it is clear to see that the three modules are closely related, and thus the final estimation is a comprehensive result that considers multiple factors.

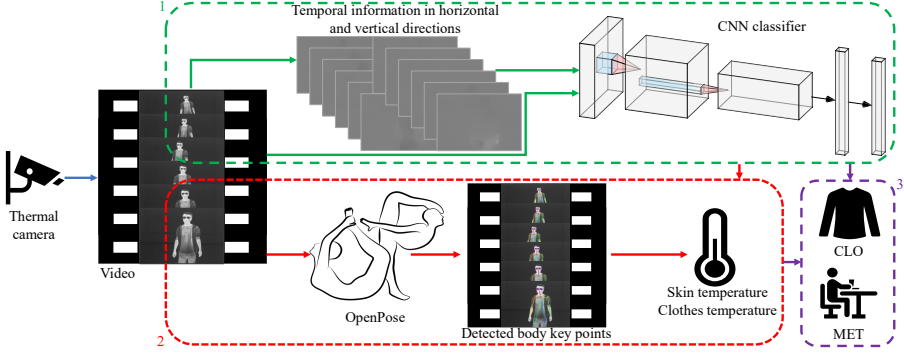


Fig. 2.8: The diagram of the work B. The green dashed box refers to the CNN module for clothes type and behavior type recognition, similar to what has been done in paper A. The red dashed box refers to the utilization of OpenPose (for detecting key body points) and the recognized clothes type to acquire a person's skin temperature and clothes temperature. The purple dashed box refers to the calculation module of I_{cl} and M that takes both the recognized type and the acquired temperatures as inputs. Image source: [4].

All the three modules have been evaluated in paper B: (1) the type recognition module realized by a CNN-based 15-category video classifier achieves an average accuracy of 95.17%, proving that the implemented spatial-temporal CNN is able to recognize not only the clothes which rely more on the spatial information but also the behaviors which rely more on the temporal information; (2) the temperature acquisition module that closely depends on the OpenPose tool works reliably, given that OpenPose can detect body key points accurately on 14123 images out of a sampled thermal dataset consisting of 14928 images (which equals an accuracy of 94.61%); (3) the estimated values of the calculation module are compared with the reference values listed in ISO standards, proving the consistency of our estimations

with the widely-accepted table look-up method.

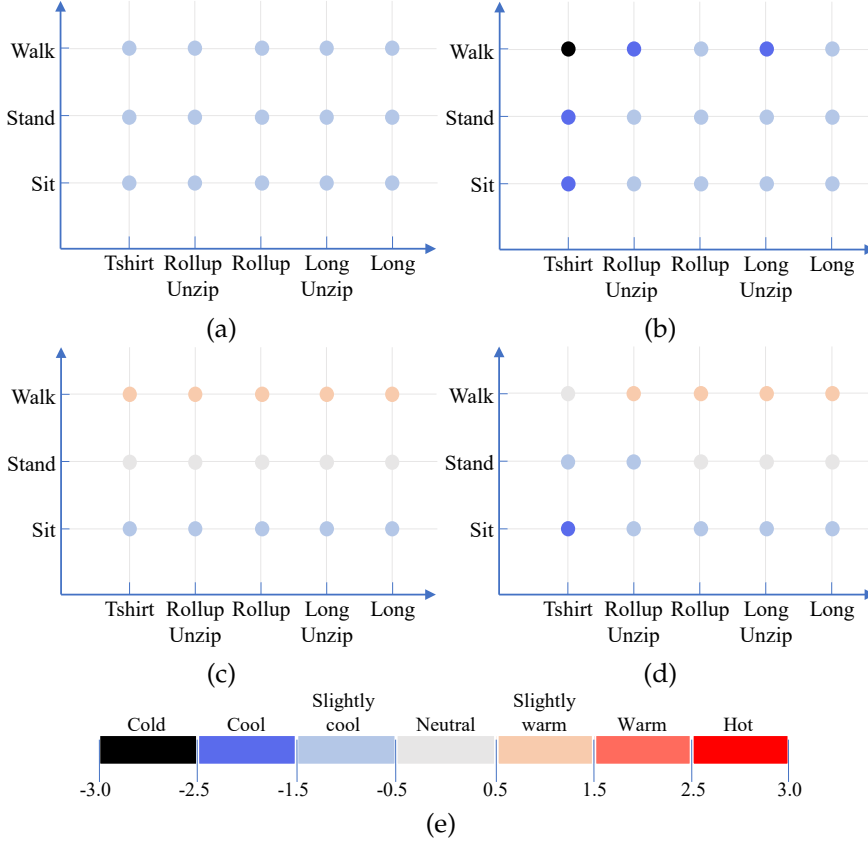


Fig. 2.9: Assessed thermal feelings from personal factors (I_{cl} and M) that are acquired from different methods. (a) Fixed I_{cl} as 0.5 clo (the representative value on warm days according to [3]) and fixed M as 55 W/m² (the value for sitting that is the dominant posture in an office). (b) Calculated I_{cl} from us and fixed M as 55 W/m². (c) Fixed I_{cl} as 0.5 clo and calculated M from us. (d) Calculated I_{cl} and M from us. (e) Seven-scale thermal feelings according to Table 2.1. Cited and adapted from [4].

Moreover, since the developed automatic estimation of I_{cl} and M is for the ultimate goal of dynamic thermal comfort assessment, by means of the CBE thermal comfort tool [10], the assessed thermal feelings of a subject are acquired based on the personal factors estimated by us and other methods. The results are illustrated in Fig. 2.9. In the figure, compared to the sub-figures from (a) to (c) where at least one factor is not estimated from the proposed method, the sub-figure (d) that uses the both factors estimated by us shows much more reasonable and dynamic changes in the assessed thermal sensation; because (d) illustrates that the person feels cool to slightly warm when

4. Contributions

his activity intensity and clothing insulation ability increase. This comparison in Fig. 2.9 further demonstrates the potential of the proposed approach in a real application.

In summary, our contributions regarding this work B are:

- We have extended the work in paper A of recognizing a person's clothes type and behavior type to the field of estimating I_{cl} and M together. As far as we know, this is the first work that estimates the both personal factors for thermal comfort assessment by use of a contactless privacy-friendly thermal camera and accompanying computer vision solutions.
- The I_{cl} estimation scheme considers multiple aspects of the sleeves status, zipper status, activity status, skin temperature, and clothes temperature, making itself a comprehensive and more accurate approach. The M estimation scheme from the recognized activity type is also proven convenient and effective.
- We have quantitatively evaluated the performance of OpenPose on thermal datasets and thus provide a helpful reference for other engineers or researchers to investigate similar research questions.

For more detailed information of this work B, please refer to paper B in Part II.

C: Clothing Insulation Rate and Metabolic Rate Estimation for Individual Thermal Comfort Assessment in Real Life [6]

Papers A and B have successfully estimated I_{cl} and M values for a single-person scenario, giving us a picture of what can be expected from a computer vision solution in *Thermal Adaptive Architecture*. On the basis of them, the extended investigation of how to employ vision-based I_{cl} and M estimation in the multiple-person dataset of a real office environment is explored in paper C.

When there is more than one person in the view captured by a camera, to achieve individual thermal comfort, each individual's thermal sensation needs to be assessed. That means each person's I_{cl} and M have to be estimated, which requires the detection and tracking of every identity. To this end, we have used a tracking-by-detection framework (incorporating a detector YOLOv5 [43] and a tracker DeepSort [44]) to track each identity across frames. At the same time, the object classification function implied in the detector YOLOv5 helps to recognize each person's clothes type and key posture since such information has been annotated in the multiple-person dataset (see Fig. 2.6).

Specifically, we decided to use the DeepSort-by-YOLOv5 framework on the multiple-person dataset due to the following reasons: (1) the researched

thermal data has much fewer details compared to normally used RGB data, therefore, utilizing these precious features/details to its utmost is of great significance; fortunately, the detector YOLOv5 has used Path Aggregation Network (PANet) [45] in its backbone, which makes the deeper layers reuse the lower-layer features in a more efficient way; (2) the tracker DeepSort has a low complexity and thus can achieve real-time performance, very suitable for our office environment with a limited number of persons instead of other crowded environments like a heavy traffic that need complex trackers; (3) in detail, DeepSort has an ability to filter out false negatives and false positives for the detection output from YOLOv5, making the integrated DeepSort-by-YOLOv5 paradigm work robustly to track each person for the further individual analysis.

Therefore, by using this tracking-by-detection framework, each person in a video will be tracked with a consistent identity (ID) across frames, like the figures shown in Fig. 2.10 where the ID numbers 1 and 2 are assigned to the two individuals over time. Then each person's I_{cl} and M can be estimated from the specific visual features of himself or herself without being interfered by other people. For I_{cl} , as what is done in paper B, the OpenPose tool for detecting key body parts, together with the recognized sleeves type from YOLOv5, can locate skin regions and clothes-covered regions. From these regions, the individual's skin temperature and clothes temperature are acquired for the further calculation of I_{cl} .



Fig. 2.10: How the tracking-by-detection module will result in for an input video. (a) and (b) are two frames from the video where ID numbers 1 and 2 indicate the two persons across frames, respectively. At the same time, each person's clothes type and posture type are recognized by the classification header of YOLOv5. Adapted from [6].

For M , we have discovered three vision-based features that can describe a person's activity intensity efficiently based on which his or her metabolic rate can be estimated. In detail, as each individual is tracked by a bounding box around, within a certain period of time, the bounding box location

4. Contributions

(represented by the center coordinates of the box) change can capture the person's overall movement, while the bounding box scale (represented by the upper-right coordinates of the box when the center of the box is set at the origin $(0, 0)$) change can capture the limb movement. Besides, the optical flow intensity from the bounding box region can capture the detailed and much smaller movements. Therefore, these three aspects of features explicitly encode a person's activity intensity, and we have proven that the higher activity intensity the person behaves, the larger bounding box changes and optical flow intensity we can extract.

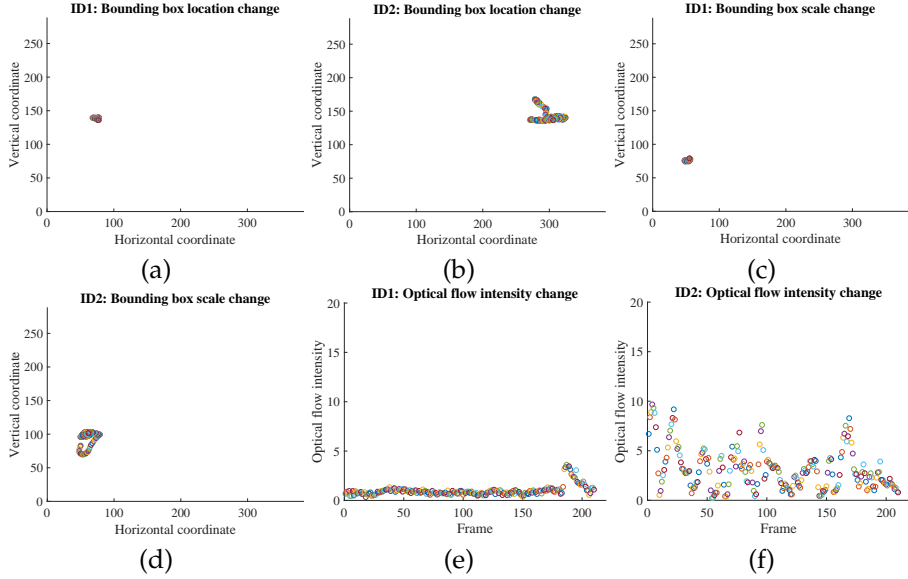


Fig. 2.11: For the ID 1 person in Fig. 2.10: the bounding box location change (a), the bounding box scale change (c), and the optical flow intensities in the bounding boxes (e). For the ID 2 person in Fig. 2.10: the bounding box location change (b), the bounding box scale change (d), and the optical flow intensities in the bounding boxes (f). Image source: [6].

As illustrated in Fig. 2.11, the ID 1 and ID 2 correspond to the two individuals in Fig. 2.10 where the ID 1 female is texting with low activity intensity while the ID 2 male is stretching arms with high activity intensity. Accordingly, take a period of 10 seconds for example, the bounding box location changes, scale changes, and the optical flow intensities of them demonstrate totally different distributions in Fig. 2.11. The points in the sub-figures Fig. 2.11(b) and (d) are much more spread out than those in (a) and (c); besides, the intensities in (f) are also much larger than those in (e), all of which are accurately describing the activity intensity difference of the ID 1 person and the ID 2 person. From this example, the link between the three vision-based features and an individual's activity intensity is verified. Therefore, a classi-

fier with the three features as inputs can be designed to categorize a person's activity intensity into one of the three levels (low, moderate, high) based on which the M value is calculated.

As a whole, we can draw the work C as the diagram in Fig. 2.12. From it, we can summarize our work in three modules:

- A tracking-by-detection module to track each individual (with a bounding box and a consistent ID across frames) and at the same time recognize the person's clothes and posture situations. This module gets a detection rate of 89.10% measured by mAP_{50} on 434 testing images and a tracking rate of 99.50% measured by Multiple Object Tracking Accuracy (MOTA) on 15 testing videos.
- A I_{cl} estimation module with both the skin temperature and clothes temperatures as its inputs that are acquired from a key body point detection tool—OpenPose. We have verified that OpenPose can detect body key points accurately on 4714 images out of a sampled thermal dataset consisting of 4901 images, which equals an accuracy of 96.18%. Besides, the estimated I_{cl} values by us have been compared with those listed in ISO 9920 [13], proving the consistency of this module and those recognized international standards.
- A M estimation module based on the categorized result of an individual's activity intensity (low, moderate, or high), realized by a classifier which feeds on the three proposed vision-based features explained above. We have verified that a random forest-based classifier can achieve a classification rate of 95.60% on a testing set of 68 samples. Besides, the estimated M values from us are very close to the reference values in the widely-used compendium of physical activities tables [18], certifying that this module is an effective and reliable method for M calculation.

With the above explanation of the proposed modules and the evaluations of them, our specific contributions regarding this work C are:

- We have initially extended the work of estimating I_{cl} and M for a single-person scenario to a multiple-person situation in real life, which provides a way to analyze each person's thermal condition in a same environment. By use of this scheme, it is possible for architecture designers to regulate the indoor microclimate that responds to different subjective thermal states.
- We have quantitatively evaluated the performance of OpenPose on a thermal dataset that has multiple persons exhibiting various types of behaviors in each frame. This investigates the potential of using OpenPose in a different mode from its initial target (RGB) and thus provides

4. Contributions

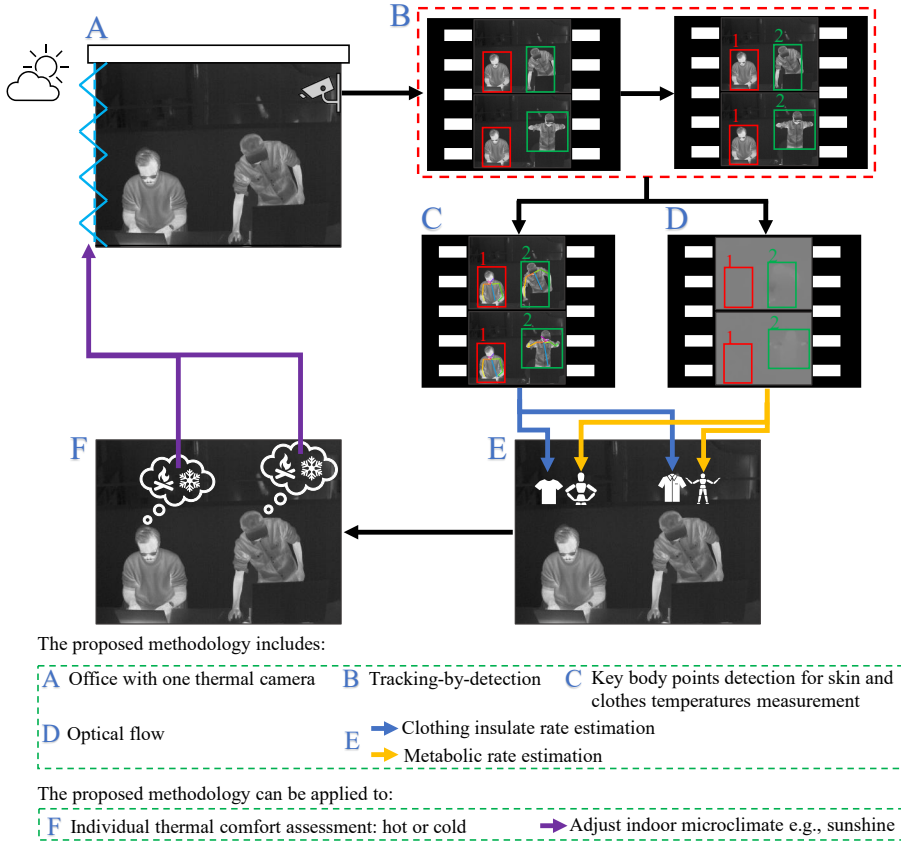


Fig. 2.12: The diagram of the work C. Both the modules that the proposed methodology includes and the applications that the methodology can be applied to are drawn. Image source: the graphical abstract for [6].

a good reference for other engineers and researchers when exploring similar research questions.

- The proposed three vision-based features are proved as good indicators of human activity intensity, which have the potential to be applied to many other problems regarding action recognition besides the estimation of metabolic rate in the thermal comfort domain.

For more detailed information of this work C, please refer to paper C in Part II.

D: Knowing Where to Look: Gender Classification in Thermal Imagery [Technical Report D in Part II]

More and more research has pointed out that females and males can have different thermal sensations in the same environment [46–53]. Therefore, if *Thermal Adaptive Architecture* can take the gender information of each occupant into consideration in addition to the estimated clothing insulate rate and metabolic rate, the applied thermal adaptive building has a potential to provide a more satisfactory microclimate for the occupants. Motivated by this, gender classification in thermal imagery has been studied.

Considering that images in thermal modality have much less information compared to the visible modality, decreased performance is anticipated for gender classification in thermal imagery. Accordingly, general approaches like designing more sophisticated networks, introducing more advanced training strategies, increasing the amount of data, etc., can be proposed to improve the performance. However, more research questions like how a data-driven model (specifically referring to a CNN in this work) obtains the ability to do gender recognition, which input region is most discriminative for a model to use to give a prediction, and if the knowledge a model has learnt for prediction is reliable, etc. are still unknown. Because existing works on gender classification in thermal imagery [54–58] just use a CNN as a “black box”.

To this end, the work D focuses on making a CNN targeted at gender classification more “transparent” by using a technique of explainable CNN—class activation mapping. This explainable AI tool can generate a heat map of the input that will show which input regions are activated for the predicted result. On top of this, a deeper understanding of gender classification can be achieved, and thus more specific strategies for the thermal modality can be introduced to lead toward a more robust performance of gender recognition in thermal imagery.

In detail, on a thermal subset (including images from 37 females and 37 males) of the Tufts facial database [59, 60], we have finetuned AlexNet [61] as a binary gender classifier. To understand which patterns/principles the CNN has learnt to make a prediction, we have used Gradient-weighted Class Activation Mapping++ (Grad-CAM++) [62] to generate the class activation maps of some training images and some validation images (thermal images of the VAP RGB-D-T dataset [63]). Examples of these activation maps are illustrated in Fig. 2.13.

From the activation maps Fig. 2.13(a), we find that a female’s hair ends are often the gender-discriminative region, which indicates the importance of hair in this task. Therefore, the corresponding regions near the ear, neck, shoulder, and upper arms that a female’s hair is expected to cover and the hair regions themselves are also activated in the validation maps Fig. 2.13(b)

4. Contributions

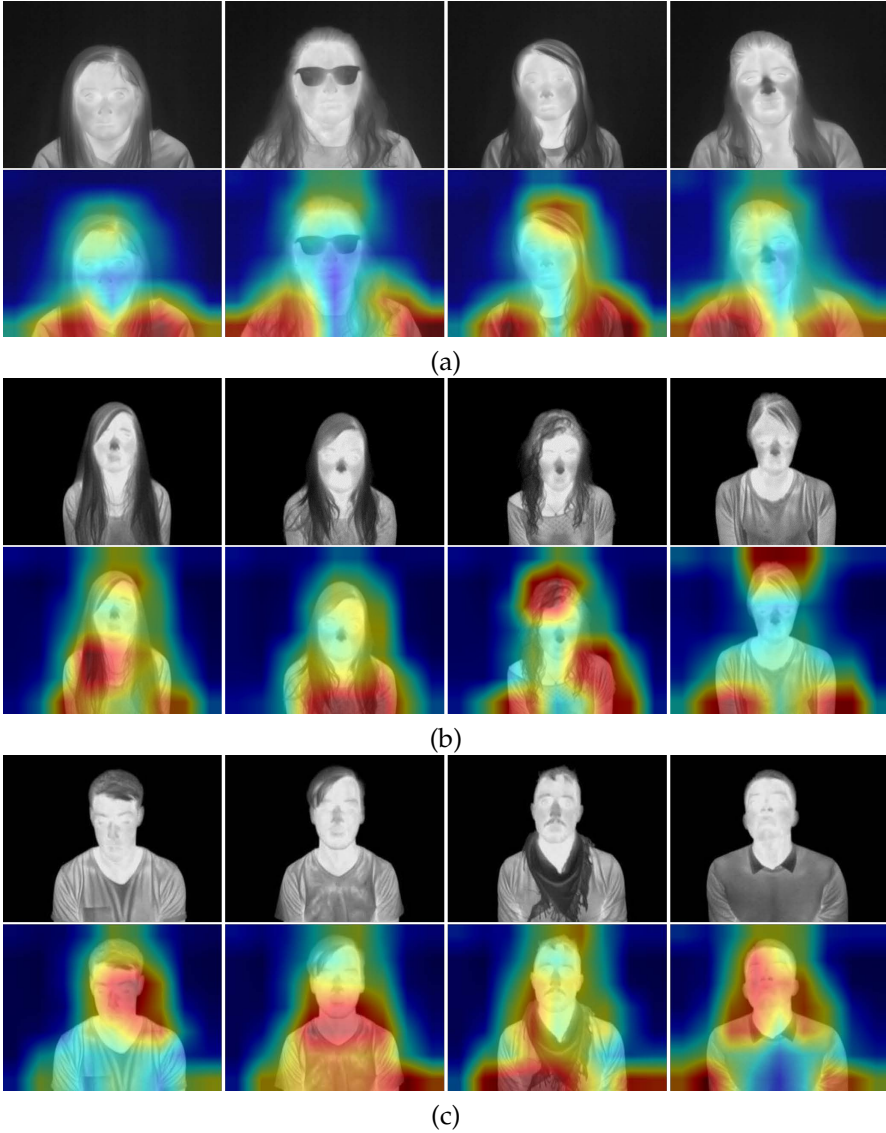


Fig. 2.13: Some sampled images from the training/validation set and their corresponding class activation maps to show what the CNN has learnt. Pay attention to the regions in red color in these maps. (a) Training images of females. (b) Validation images of females. (c) Validation images of males. Image source: technical report D in Part II.

and (c), based on which the CNN distinguishes a female from a male. We attribute this phenomenon to an observation that females are expected to have shoulder-length or longer hair than males. However, this observation result

concluded by the classifier is potentially risky, as hair features are extrinsic that have no ability to describe an identity intrinsically. Hence, the finetuned CNN will be very prone to data imbalance problems and may be problematic in the application phase.

In fact, the activated regions in Fig. 2.13 have shown that the CNN has been influenced by the data bias problem, as we further find that there are more females with shoulder-length and long hair than those with short hair in the training set. To further explore how this bias influences the gender-discriminative region and whether it is possible to extract intrinsic features for gender classification in thermal imagery, we have done two extended experiments: (1) Exp1: the training set only includes females with long hair and hijabs similar to long hair; (2) Exp2: the training set only includes females with short hair and ponytails. The hypothesis is that Exp1 will mainly focus on the long hair area to make a prediction and hence neglect other features, which will mis-classify a female with short hair as a male; Exp2 can take more information into consideration and thus potentially achieve a more robust classification result.

After the same finetuning phase as the basic experiment, on the same validation set, both the CNNs of Exp1 and Exp2 perform as hypothesized, and the validation classification rate increases from 80% (Exp1) to 96.25% (Exp2). For better illustration, the class activation maps of some validation images from Grad-CAM++ are shown in Fig. 2.14 including both maps corresponding to Exp1 and Exp2. Additional information is that these maps in the figure are an assembled version obtained by a person-wise average; this assembling process considers all the activation maps of a same person as each person for validation is captured several times; hence, the maps in Fig. 2.14(b) and (c) gain the capacity to comprehensively represent all the activated regions in Exp1 and Exp2. From Fig. 2.14(b) and (c), it can be observed that from Exp1 to Exp2, the gender-discriminative regions are updated from areas indicating a shoulder-length and longer hair to the top of the head and more general broader areas. However, these results still imply the importance of the hairstyle difference of genders for such a task, and more importance intrinsic facial features seems not to play a major role.

Together with other extensive analyses in the technical report D, it is concluded that a finetuned CNN on the Tufts dataset for recognizing gender in thermal imagery is more focused on hair features rather than the facial features. This makes sense considering that the detailed facial information is not acquired from a thermal camera, but one problem is that these non-facial features are prone to training data biases which will cause the extraction of improper features that lead to degraded performance. In contrast, though RGB visible datasets also suffer from biases, we believe that the much more available details in an RGB image can provide sufficient information based on which the impact of a bias can be mitigated to some degree.

4. Contributions

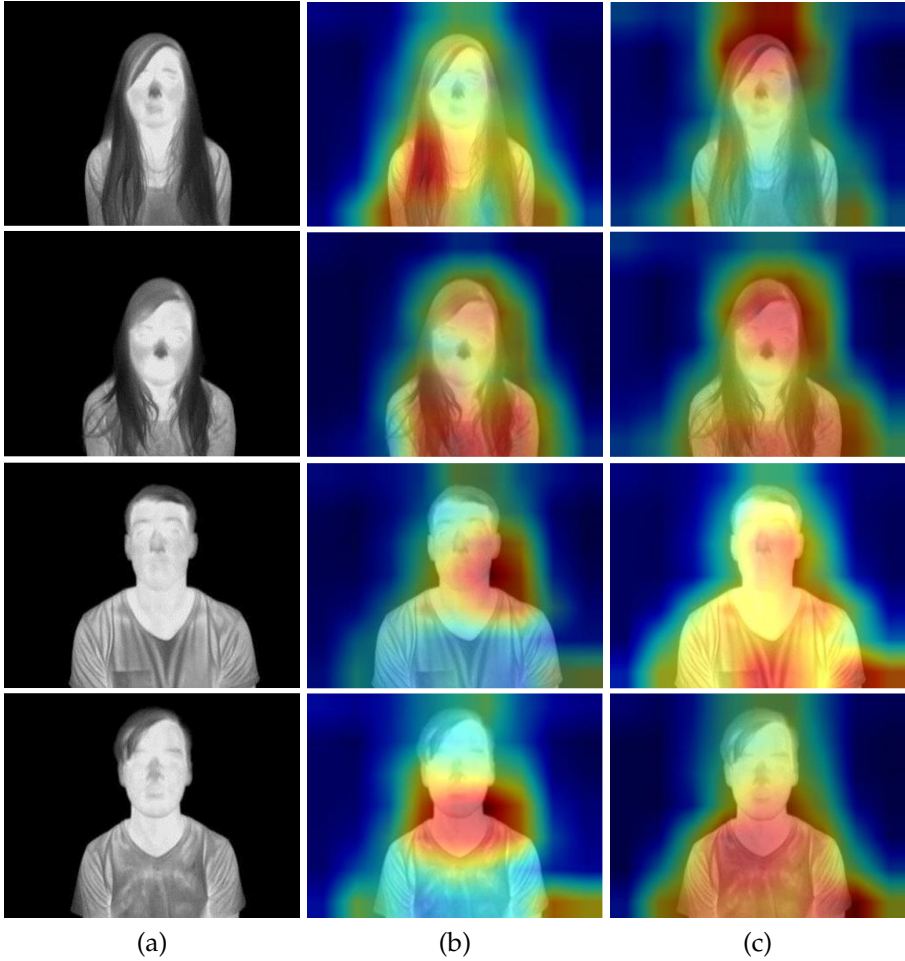


Fig. 2.14: Example validation images and their assembled class activation maps. (a) A participant’s frontal image. (b) Exp1-generated assembled class activation maps. (c) Exp2-generated assembled class activation maps. Image source: technical report D in Part II.

We further suggest some good empirical practices to find biases in thermal imagery more effectively—keep an eye on the object/phenomenon that has thermal/temperature properties, for instance, the data collection duration that affects the environment temperature, a warm-blooded animal that exhibits a similar temperature distribution as a person, glasses/hats worn by persons but exhibit a totally distinct temperature distribution from humans, etc. Besides, a CNN overfitting the training set can extract every detailed feature of the training images. And thus, using the trained model to generate class activation maps for the training images can visualize what the CNN has

learnt to illustrate if a bias exists.

As a whole, our specific contributions regarding this technical report D are:

- We have done a visualization-based analysis to find gender-discriminative regions for the CNN-based gender classification in thermal imagery. Accordingly, we have discovered the importance of extrinsic features for this task, which gives a “transparent” understanding of how a CNN recognizes a female or a male.
- Due to the concern that extrinsic features are inclined to be influenced by training data biases, we have also done some extended experiments to investigate the influence of hairstyle stereotypes on the performance and accordingly provide good recommendations to find data imbalances for more robust training and testing.

For more detailed information of this work D, please refer to the technical report D in Part II.

Sub-conclusion

These works A, B, C, and D were done along with the progress of the PhD project. Among them, the first three works on I_{cl} and M estimation follow the sequence from easy to hard, from a single-person scenario to a multiple-person scenario, from a laboratory study to a real-life field study, and from a video-level recognition to a frame-by-frame recognition. The last work on gender classification is ongoing work inspired by the newly-learnt knowledge of the gender difference in thermal comfort assessment. All the works together emphasize the importance of the sleeves status for I_{cl} , OpenPose for I_{cl} , the key posture for M , the optical flow for M , the tracking function, and the gender classification for an “individual” assessment.

References

- [1] World Health Organization, “World health organization basic documents,” https://apps.who.int/gb/bd/pdf_files/BD_49th-en.pdf, 2020, last accessed: March, 2022.
- [2] U.S. Energy Information Administration, “2018 commercial buildings energy consumption survey data,” <https://www.eia.gov/consumption/commercial/data/2018/>, last accessed: March, 2022.
- [3] Refrigerating The American Society of Heating and Air-Conditioning Engineers (ASHRAE), “Ashrae standard: Thermal environmental conditions for human occupancy,” [http:](http://)

References

- //arco-hvac.ir/wp-content/uploads/2015/11/ASHRAE-55-2010.pdf, last accessed: March, 2022.
- [4] Jinsong Liu, Isak Worre Foged, and Thomas B Moeslund, "Automatic estimation of clothing insulation rate and metabolic rate for dynamic thermal comfort assessment," *Pattern Analysis and Applications*, pp. 1–16, 2021.
 - [5] Povl Ove Fanger, "Assessment of man's thermal comfort in practice," *Occupational and Environmental Medicine*, vol. 30, no. 4, pp. 313–324, 1973.
 - [6] Jinsong Liu, Isak Worre Foged, and Thomas B Moeslund, "Clothing insulation rate and metabolic rate estimation for individual thermal comfort assessment in real life," *Sensors*, vol. 22, no. 2, pp. 619, 2022.
 - [7] P.O. Fanger, *Thermal Comfort: Analysis and Applications in Environmental Engineering*, McGraw-Hill, 1970.
 - [8] International Organization for Standardization, "Ergonomics of the thermal environment — analytical determination and interpretation of thermal comfort using calculation of the pmv and ppd indices and local thermal comfort criteria," <https://www.sis.se/api/document/preview/907006/>, last accessed: March, 2022.
 - [9] Federico Tartarini, Stefano Schiavon, Toby Cheung, and Tyler Hoyt, "Cbe thermal comfort tool: Online tool for thermal comfort calculations and visualizations," *SoftwareX*, vol. 12, pp. 100563, 2020.
 - [10] Federico Tartarini, Stefano Schiavon, Toby Cheung, and Tyler Hoyt, "Cbe thermal comfort tool," <https://comfort.cbe.berkeley.edu/>, 2020, last accessed: March, 2022.
 - [11] International Organization for Standardization, "Ergonomics of the thermal environment — instruments for measuring physical quantities," <https://www.sis.se/api/document/preview/615884/>, last accessed: March, 2022.
 - [12] International Organization for Standardization, "Ergonomics of the thermal environment — analytical determination and interpretation of heat stress using calculation of the predicted heat strain," <https://www.iso.org/standard/37600.html/>, last accessed: March, 2022.
 - [13] International Organization for Standardization, "Ergonomics of the thermal environment — estimation of thermal insulation and water vapour resistance of a clothing ensemble," <https://www.iso.org/standard/39257.html/>, last accessed: March, 2022.

- [14] International Organization for Standardization, "Ergonomics of the thermal environment — determination of metabolic rate," <https://www.iso.org/standard/34251.html/>, last accessed: March, 2022.
- [15] Jakob Petersson and Amitava Halder, "Updated database of clothing thermal insulation and vapor permeability values of western ensembles for use in ashrae standard 55, iso 7730, and iso 9920," *ASHRAE Transactions*, vol. 127, pp. 773–799, 2021.
- [16] Yin Tang, Zixiong Su, Hang Yu, Kege Zhang, Chaoen Li, and Hai Ye, "A database of clothing overall and local insulation and prediction models for estimating ensembles' insulation," *Building and Environment*, vol. 207, pp. 108418, 2022.
- [17] Barbara E Ainsworth, William L Haskell, Melicia C Whitt, Melinda L Irwin, Ann M Swartz, Scott J Strath, WILLIAM L O'Brien, David R Bassett, Kathryn H Schmitz, Patricia O Emplaineourt, et al., "Compendium of physical activities: an update of activity codes and met intensities," *Medicine and science in sports and exercise*, vol. 32, no. 9; SUPP/1, pp. S498–S504, 2000.
- [18] Barbara E Ainsworth, William L Haskell, Stephen D Herrmann, Nathanael Meckes, David R Bassett, Catrine Tudor-Locke, Jennifer L Greer, Jesse Vezina, Melicia C Whitt-Glover, and Arthur S Leon, "2011 compendium of physical activities: a second update of codes and met values," *Med Sci Sports Exerc*, vol. 43, no. 8, pp. 1575–1581, 2011.
- [19] Jack Ngarambe, Geun Young Yun, and Gon Kim, "Prediction of indoor clothing insulation levels: A deep learning approach," *Energy and Buildings*, vol. 202, pp. 109402, 2019.
- [20] Michele De Carli, Bjarne W Olesen, Angelo Zarrella, and Roberto Zecchin, "People's clothing behaviour according to external weather and indoor environment," *Building and Environment*, vol. 42, no. 12, pp. 3965–3973, 2007.
- [21] Hiroki Matsumoto, Yoshio Iwai, and Hiroshi Ishiguro, "Estimation of thermal comfort by measuring clo value without contact," in *MVA*. Cite-seer, 2011, pp. 491–494.
- [22] Siliang Lu and Erica Cochran Hameen, "Integrated ir vision sensor for online clothing insulation measurement," in *Proceedings of the 23rd CAADRIA Conference*, 2018, pp. 565–573.
- [23] Mohammad H Hasan, Fadi Alsaleem, and Mostafa Rafaie, "Sensitivity study for the pmv thermal comfort model and the use of wearable

- devices biometric data for metabolic rate estimation," *Building and Environment*, vol. 110, pp. 173–183, 2016.
- [24] Andrea Calvaresi, Marco Arnesano, Filippo Pietroni, and Gian Marco Revel, "Measuring metabolic rate to improve comfort management in buildings," *Environmental Engineering & Management Journal (EEMJ)*, vol. 17, no. 10, 2018.
- [25] Fitbit InC, "Fitbit charge hr: Product manual," https://help.fitbit.com/manuals/manual_charge_hr_en_US.pdf, last accessed: March, 2022.
- [26] Zephyr Technology, "Bioharness 3.0: User manual," <https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf>, last accessed: March, 2022.
- [27] COSMED, "K5: The one choice metabolic system for both laboratory and field testing," <https://www.cosmed.com/en/products/cardio-pulmonary-exercise-test/k5>, last accessed: March, 2022.
- [28] Yongchao Zhai, Minghui Li, Siru Gao, Liu Yang, Hui Zhang, Edward Arens, and Yunfei Gao, "Indirect calorimetry on the metabolic rate of sitting, standing and walking office activities," *Building and Environment*, vol. 145, pp. 77–84, 2018.
- [29] East Medic Corporation, "Ae-100i," http://www.east-medic.jp/rm_sm_im/ae-100i/, last accessed: March, 2022.
- [30] Akihisa Nomoto, Ryo Hisayama, Shu Yoda, Mizuho Akimoto, Masayuki Ogata, Hitomi Tsutsumi, and Shin-ichi Tanabe, "Indirect calorimetry of metabolic rate in college-age japanese subjects during various office activities," *Building and Environment*, vol. 199, pp. 107909, 2021.
- [31] Jeong-Hoon Lee, Young-Keun Kim, Kyung-Soo Kim, and Soohyun Kim, "Estimating clothing thermal insulation using an infrared camera," *Sensors*, vol. 16, no. 3, pp. 341, 2016.
- [32] Tanveer Ahmad, Taimur Rashid, Hassan Abbas Khawaja, and Mojtaba Moatamedi, "Study of the required thermal insulation (ireq) of clothing using infrared imaging," *The International Journal of Multiphysics*, vol. 11, no. 4, 2017.
- [33] Kyungsoo Lee, Haneul Choi, Hyungkeun Kim, Daeung Danny Kim, and Taeyeon Kim, "Assessment of a real-time prediction method for high clothing thermal insulation using a thermoregulation model and an infrared camera," *Atmosphere*, vol. 11, no. 1, pp. 106, 2020.

- [34] Haneul Choi, HooSeung Na, Taehung Kim, and Taeyeon Kim, "Vision-based estimation of clothing insulation for building control: A case study of residential buildings," *Building and Environment*, vol. 202, pp. 108036, 2021.
- [35] HooSeung Na and Taeyeon Kim, "Development of metabolic rate prediction model using deep learning via kinect camera in an indoor environment," in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, 2019, vol. 609, p. 042036.
- [36] HooSeung Na, Joon-Ho Choi, HoSeong Kim, and Taeyeon Kim, "Development of a human metabolic rate prediction model based on the use of kinect-camera generated visual data-driven approaches," *Building and Environment*, vol. 160, pp. 106216, 2019.
- [37] Jinsong Liu, Isak Worre Foged, and Thomas B Moeslund, "Vision-based individual factors acquisition for thermal comfort assessment in a built environment," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 662–666.
- [38] Xenics, "Smart and affordable xenics gobi-384," <https://www.xenics.com/smart-and-affordable-xenics-gobi-384/>, last accessed: March, 2022.
- [39] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.
- [40] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Asian conference on computer vision*. Springer, 2018, pp. 363–378.
- [41] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.
- [42] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.
- [43] Glenn Jocher et. al., "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," <https://doi.org/10.5281/zenodo.6222936>, Feb. 2022.

- [44] Nicolai Wojke, Alex Bewley, and Dietrich Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.
- [45] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.
- [46] Sami Karjalainen, "Thermal comfort and gender: a literature review," *Indoor air*, vol. 22, no. 2, pp. 96–109, 2012.
- [47] Jinhua Hu, Yingdong He, Xiaoli Hao, Nianping Li, Yuan Su, and Huaidei Qu, "Optimal temperature ranges considering gender differences in thermal comfort, work performance, and sick building syndrome: A winter field study in university classrooms," *Energy and Buildings*, vol. 254, pp. 111554, 2022.
- [48] Jéssica Kuntz Maykot, Ricardo Forgiarini Rupp, and Eneide Ghisi, "Assessment of gender on requirements for thermal comfort in office buildings located in the brazilian humid subtropical climate," *Energy and Buildings*, vol. 158, pp. 1170–1183, 2018.
- [49] Jéssica Kuntz Maykot, Ricardo Forgiarini Rupp, and Eneide Ghisi, "A field study about gender and thermal comfort temperatures in office buildings," *Energy and Buildings*, vol. 178, pp. 254–264, 2018.
- [50] Madhavi Indraganti, "Gender differences in thermal comfort and satisfaction in offices in gcc and asia," in *Gulf conference on sustainable built environment*. Springer, 2020, pp. 483–497.
- [51] Thomas Parkinson, Stefano Schiavon, Richard de Dear, and Gail Brager, "Overcooling of offices reveals gender inequity in thermal comfort," *Scientific reports*, vol. 11, no. 1, pp. 1–7, 2021.
- [52] Mina Jowkar, Hom Bahadur Rijal, Azadeh Montazami, James Brusey, and Alenka Temeljotov-Salaj, "The influence of acclimatization, age and gender-related differences on thermal perception in university buildings: Case studies in scotland and england," *Building and Environment*, vol. 179, pp. 106933, 2020.
- [53] Madhavi Indraganti and Michael A Humphreys, "A comparative study of gender differences in thermal comfort and environmental satisfaction in air-conditioned offices in qatar, india, and japan," *Building and Environment*, vol. 206, pp. 108297, 2021.
- [54] Dat Tien Nguyen and Kang Ryoung Park, "Body-based gender recognition using images from visible and thermal cameras," *Sensors*, vol. 16, no. 2, pp. 156, 2016.

- [55] Dat Tien Nguyen, Ki Wan Kim, Hyung Gil Hong, Ja Hyung Koo, Min Cheol Kim, and Kang Ryoung Park, "Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction," *Sensors*, vol. 17, no. 3, pp. 637, 2017.
- [56] Shangfei Wang, Zhen Gao, Shan He, Menghua He, and Qiang Ji, "Gender recognition from visible and thermal infrared facial images," *Multimedia Tools and Applications*, vol. 75, no. 14, pp. 8419–8442, 2016.
- [57] Muhammad Ali Farooq, Hossein Javidnia, and Peter Corcoran, "Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system," *Journal of Electronic Imaging*, vol. 29, no. 6, pp. 063004, 2020.
- [58] Kateřina Přihodová and Jakub Jech, "Gender recognition using thermal images from uav," in *2021 International Conference on Information and Digital Technologies (IDT)*. IEEE, 2021, pp. 83–88.
- [59] Sensing The Panetta Visualization and Simulation Research Laboratory, "The tufts face database," <http://tdface.ece.tufts.edu/>, last accessed: April, 2022.
- [60] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A Taylor, Arash Samani, et al., "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.
- [61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [62] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.
- [63] Olegs Nikisins, Kamal Nasrollahi, Modris Greitans, and Thomas B Moeslund, "Rgb-dt based face recognition," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 1716–1721.

Chapter 3

Safe Harbor

1 Background

Over the past decades, we humans have witnessed a great development in various social sectors including food supply, education, health care, etc., making raiment and daily bread not the minimum requirement in our life. Accordingly, how to live a life with more happiness has aroused growing attention, among which security is the most important factor to guarantee other expectations.

As a result, surveillance cameras have been installed in more and more public and private places to provide safety as all-around as possible. To operate the existing surveillance systems, many agencies rely on human resources to provide manual monitoring by looking at the captured videos for 24 hours a day, which is inefficient and expensive, not to mention the increasing error rate when the operator becomes tired after a concentrated focus for long periods of time.

Hence, there is a pressing need for an automated surveillance system to assist humans, not only to improve the efficiency for lower consumption of manpower but also to reduce human mistakes for better stability. The development of such a smart surveillance tool is the transformation from the traditional function of video display to the capability of giving analyzed results like a human being, so that beforehand warnings, the immediate rescue of accidents, evidence search afterward in helping solve crimes, and others are possible.

To realize this, besides the simple acquisition-display pipeline, an autonomous surveillance framework must integrate new function modules which are generally data preprocessing, data analysis, decision making, etc. The preprocessing of videos occurs when there is a need to improve the spatial resolution, remove noises like blurring effects, enhance the contrast

resulting from bad illumination conditions, or other situations. This pre-processing makes the data in a better format for subsequent analysis and decision making. Nowadays, these two modules are benefited from the mainstream data-driven computer vision algorithms that leverage vast amounts of data (in a similar way that a human gains precious experience from years of manual observation on displayed real-time videos) to decide whether a scene it looks at needs further involvement of human operators.

In this way, an autonomous surveillance system using multiple cameras in the same mode (e.g., visible RGB), fixed camera(s), moving camera(s), or multi-modal cameras (e.g., visible RGB, thermal, depth) can be applied to many applications in various indoor and outdoor scenarios like bus terminals, airports, department stores, hospitals, highways, etc. [1–4]. This trend makes it possible to realize the detection and tracking of pedestrians for safer traffic [5], fall detection for older people to prevent serious health issues [6], fire detection to reduce injuries and deaths [7], and other specific purposes.

Besides, the feeling of not being invaded of privacy is as important as the feeling of safety, especially for some people who feel uncomfortable when they are under monitoring. For this concern, the widespread surveillance cameras have to obey regulations like GDPR to protect the facial and other personal information of users so that a good balance between privacy-preserving and safety guarantees can be achieved. Fortunately, thermal cameras provide a perfect solution for this condition as mentioned in Chapter 1. Therefore, within the context of an automated surveillance system, the improved efficiency, decreased manual failures, and preserved personal information will together set a better living standard for all people.

2 Introduction

In this particular PhD project *Safe Harbor*, the location under monitoring by a surveillance system is a harbor area in Aalborg, a coastal city in Denmark, and the specific region under surveillance is shown in Fig. 3.1(a). There are two fixed cameras whose field of view (FOV) are 22° and 11° , respectively, to monitor the harbor front area. There is another Pan-Tilt-Zoom (PTZ) camera with its FOV of 11° monitoring the water area for manually searching for objects in the water. This PhD project is specifically using the 22° fixed camera to analyze human activities in the harbor front, and an image sample captured by this camera is shown in Fig. 3.1(b).

The choice of monitoring such a harbor front is on the basis of its important function in people's lives, as it provides the site not only for daily traffic routes but also for leisure hours spent. Therefore, the surveillance of this harbor front can provide a wealth of information that involves monitoring traffic situations, statistics study of outdoor activities, observing social distance on

2. Introduction



Fig. 3.1: The specific region under monitoring by the surveillance system with three cameras is shown in (a). An thermal image acquired by the camera with a FOV of 22° is shown in (b). Image source: (a) [8].

coronavirus days, preventing and rescuing drowning accidents, and many other aspects.

Among these things, some are more significant as they may involve a higher probability of danger or harm. That is, traffic situations referring in particular to jams and even accidents due to the narrower path along the harbor front compared to other normal roads, increased epidemic spreading from people gathering without considering the social distance, and potential drowning when a person sits or stands near the harbor edge especially when there is no witness nearby. Some of these situations are shown in Fig. 3.2.



Fig. 3.2: Examples of harbor front scenes. The left figure (a) shows heavy traffic flows and people crowds. The right figure (b) shows a potential danger of sitting near the harbor edge. The green line in the figure locates the harbor edge, the left of which is the water area.

On the one hand, people may think these dangerous events are far from their lives and thus do not always pay attention to them. However, taking

drowning accidents as an example, “there is an estimated number of 372000 people died from drowning in 2012, and hence drowning has been the third leading unintentional injury killer in the world”, according to the report from the World Health Organization (WHO) [9]. Solely in Denmark, “1565 people drowned in the years from 2001 to 2014, and 390 (25%) of these deaths occurred at harbor areas” [8]. Therefore, extra attention is absolutely necessary to these issues to reduce injuries and deaths.

On the other hand, compared with normal events which account for the overwhelming majority, the events of danger or harms are extremely rare and thus can be treated as anomalies. Therefore, if there is a need for an automated surveillance system to assist humans in finding such accidents effectively, following the commonly-used data-driven computer vision strategy that uses a large amount of annotated data to teach a CNN what the dangerous scenes look like will be extremely difficult. Besides, because there are a lot of different kinds of events requiring extra human control or rescue, it will be impossible to include all the event types through data collection.

In summary, *Safe Harbor* aims at monitoring a harbor front region and detecting anomalies of emergencies and potentially dangerous incidents that need extra attention or even immediate controls and rescues by professionals, by the use of a fixed thermal camera as hardware and computer vision solutions that can overcome the above-mentioned challenges as software. This means that *Safe Harbor* is a project that defines a relatively large range of research questions as long as they help to improve the safety level in the harbor front area. To specify the research area, we have aroused two questions to be investigated—a specific one on drowning accident prevention and a general one on anomaly detection on thermal data.

3 Related Work

To chime in with the defined research directions of drowning accident prevention and anomaly detection on thermal data, the representative work related to the two topics will be introduced. For the avoidance of redundancy, other related studies have been introduced in papers E, F, G, and H appended in Part III.

3.1 Drowning Accident Prevention

In the context of a (semi-)automatic surveillance system, studies on drowning prevention are surprisingly rare. For a clear description, we have categorized these very limited number of studies into two types according to their application areas.

3. Related Work

The first type is in the context of swimming in pools equipped with underwater and above-water cameras instead of falling into water in our case. [10–13] are such representatives that define a set of rules to describe a drowning swimmer: moving very slowly or staying in a small water region, abnormal limbs movements like irregularly and rapidly tapping the water with arms, underwater more than a predefined time threshold, etc. To compare a swimmer to these rules, the swimmer has to be detected in the water by a human detector or background segmentation and then analyzed to get his or her velocity and posture. If a swimmer exhibits these drowning signs, alarms will be sent out to lifeguards or authorities to provide an immediate rescue. Another realization of drowning detection [14] is more direct from the perspective of CNN-based computer vision algorithms. It first collects a dataset of images demonstrating drowning poses and non-drowning poses, based on which a binary classifier on top of AlexNet [15] is trained to predict whether a testing frame of a person is drowning or not.

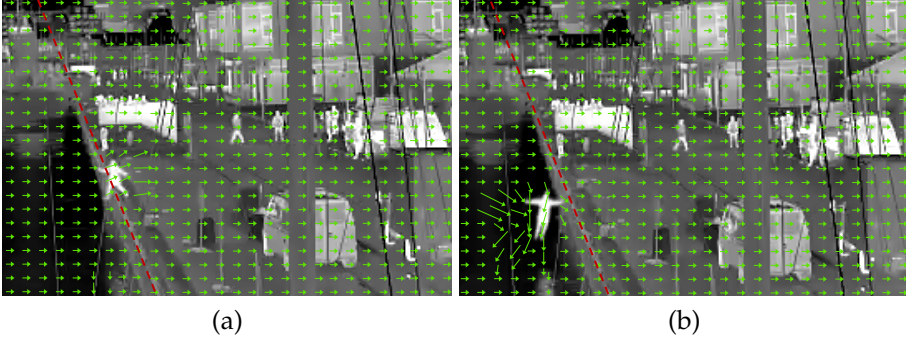


Fig. 3.3: Illustrations of the optical flow (green arrows) when a person starts falling into water from the harbor edge (a) and when the person is just before hitting the water (b). The red dashed line locates the harbor edge, the left of which is the water area. Both images are adapted from [16] for better visualization.

The second type is in the context of falling into water near the sea, lake, or pool. However, drowning prevention designed for these areas is extremely rare. A representative study [16] focusing on the same harbor front with us realizes the prevention of drowning accidents by calculating the human movement in the form of optical flow magnitude. If the magnitude in the water area is larger than a predefined threshold, there is a just happening falling accident from the harbor to the water, which indicates an early drowning. The subsequent alarm can be aroused to authorities very quickly before the person hits the water. The feasibility of this idea is visualized in Fig. 3.3. When there is no falling accident or when a person is just leaving the harbor edge to the water, the optical flow pattern in the water area is weak. However, when a person is falling and just before hitting the water, the optical

flow pattern in the water area will fluctuate significantly with a much larger magnitude that predicts a falling accident.

3.2 Anomaly Detection

The incidents that need extra attention from professionals to provide controls and rescues are much rarer compared to overwhelmingly frequent normal events. This property determines that the detection of aforesaid scarce incidents can be effectively realized by anomaly detection, according to its definition that “anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior” [17].

In terms of the popular deep learning guided computer vision, supervised classification through a large number of labelled anomalies is unrealistic, and thus unsupervised/self-supervised algorithms without labeling are more preferred. [18] initially uses a convolutional autoencoder (AE) [19] which is trained with only normal patterns to reconstruct the input; because the AE has not seen anomalies in the training phase, the reconstructed output will be more similar to the normality no matter what input is fed into the AE. Hence, the difference between an anomalous input and its output will be much larger than that of a normal input, based on which an anomaly is detected. It is worth mentioning that the brief description of this study is the typical pipeline of AE-based anomaly detection in today’s computer vision solutions. On the strength of [18], research [20] considers the temporal information across frames of a video volume in detecting anomalies by incorporating convolutional Long Short Term Memory (LSTM) models [21] in both the encoder and the decoder of the AE.

In the wake of the extensive application of this AE-reconstruction-error paradigm, people have found an undesirable problem—sometimes the capacity of the deep autoencoder is too powerful and thus it can reconstruct an anomaly quite well, which will cause some missing detections. To amend this, two appreciated improvements are invented. (1) [22] proposes a future-frame-prediction paradigm that predicts the next frame based on its historical frames by using a U-shape network [23] as the generator, making the prediction decoupled with the current frame. The feasibility of this method is on the basis that if the generator has been trained to predict the next normal pattern successfully, a normal frame is predictable while an abnormal one is unpredictable in the testing phase. (2) [24] proposes a memory-augmented deep AE (MemAE). This MemAE has a memory slot to store prototypical features of normal data at the training stage. Then at the testing stage, the encoded features of any input are not directly fed to the decoder to reconstruct the input. Instead, these features are used to retrieve the most relevant stored features from the memory slot as the “food” to the decoder, forming a new pipeline that does not strictly rely on the testing input. An illustration

of MemAE is shown in Fig. 3.4 that helps to give an intuitive explanation.

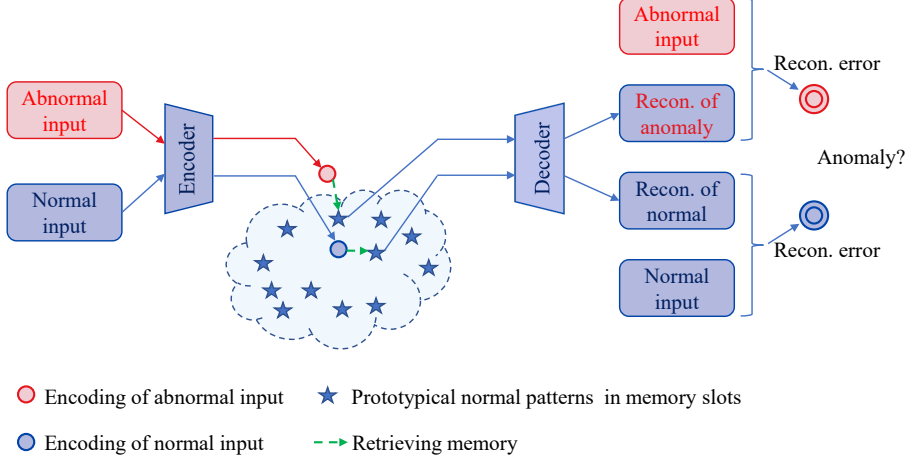


Fig. 3.4: The diagram of anomaly detection using MemAE. “Recon.” is short for reconstruction. With MemAE, the reconstruction error of an anomaly input should be always larger than that of a normal input. For the visualization simplicity, [24] assumes only one retrieved feature is required. Image source: [24].

When further expanding the scope of AE-based anomaly detection, people have found that a well-trained AE on one scene cannot be directly applied to another scene. For example, an AE designed on ShanghaiTech dataset [25] needs the retraining of the AE if it has to be used on Avenue dataset [26]. To save the inconvenience for a cross-domain application provided that the anomalies are defined identically in both scenarios, studies [27, 28] propose object-centric AEs that only take the object of interest instead of the full image as the input. For this, an object detector like SSD [29] or YOLOv3 [30] has to be implemented in the first stage of the pipeline for frame-by-frame object detection. This object-centric idea calculates the final anomaly score of a frame from the anomaly scores of all the detected objects within it, and here all of the scores are regarding a pixel-level difference between the reconstructed result and the input.

As a whole, all of these studies are on visible data, that is, Avenue [26], ShanghaiTech [25], UCSD [31], UMN [32], Subway [33], etc. Anomaly detection on thermal data is enormously rare as far as we have surveyed, making itself an open research question to be explored.

4 Contributions

Even though existing studies have been applied in preventing drowning accidents and detecting anomalies, the specific research on how to do them

in the area of harbor fronts, especially in thermal imagery as posted in *Safe Harbor* is scarce. The main differences between the available work and our realizations are listed in Table 3.1.

Table 3.1: The main differences between the available work and our realizations in preventing drowning accidents and detecting anomalies.

Drowning prevention		Anomaly detection	
Others	Ours	Others	Ours
Drowning detection in the water.	Drowning <u>warning in advance</u> if a person is very near the water.	Evaluation on visible data.	Evaluation on <u>new thermal data</u> we have collected.
Very rare data of people falling into water or in the water.	Collected a dataset by using a <u>human-like doll</u> to simulate a person falling into water to <u>reduce injuries</u> .	Evaluation on data with short time duration.	Evaluation on <u>long-term data</u> for a <u>real-world</u> application.

In the table, the keywords of our contributions on *Safe Harbor* during the PhD study are underlined with wavy lines. The following contents will use the studies done in papers E, F, G, and H in Part III to explain these points in detail.

4.1 Data Collection

As listed in Table 3.1, there is very scarce data on people falling into water or in the water for the research on drowning prevention. Besides, the benchmark datasets (Avenue [26], ShanghaiTech [25], UCSD [31], UMN [32], Subway [33]) for anomaly detection have a common problem—the duration is short (only several minutes or hours). The algorithms evaluated on them, therefore, might be problematic for a long-term surveillance system that runs for months and years in the real world.

To settle these matters, during the PhD study we have collected a new falling-into-water dataset and a long-term thermal dataset spanning eight months. Both of them were captured by the fixed camera of 22° FOV which is actually a thermal-visible bi-spectrum camera [34], but only the thermal channel is used in *Safe Harbor* considering pedestrians’ privacy concerns.

Falling-into-water Dataset as Part of Paper H [35]

If a falling from the harbor edge can be detected before the person hits the water, there will be a longer response time for lifeguards or other professionals to provide assistance, which will have a higher rescue rate than the situation where the person is detected in the water.

Therefore, a falling-into-water dataset is so valuable that we accordingly gathered such a dataset in November 2021 and then made it public as part of paper H—*Imitating Emergencies: Generating Thermal Surveillance Fall Data Using Low-Cost Human-like Dolls* [35].

The main innovation of this dataset paper is using a low-cost inflatable doll that is similar to a real person to simulate falling incidents from a harbor front to the water. This idea not only reduces the potential injuries if human volunteers are involved in such a dangerous scenario but also avoids the problem of unconvincing simulation if we use computer software to generate artificial data.

During the dataset construction phase, to make the doll visualized as a human-like appearance in thermal imagery, it must have a stable and higher temperature than the environment. However, the doll has no ability to generate heat from itself like us humans. We, therefore, solved this by dressing the doll in clothes on which hot water was poured to keep the doll warm enough. A picture taken indoors of the clothed doll is in Fig. 3.5(a). We chose this heating method instead of chemical heating pads and electrical warm vests. Because when the doll is in the water, the chemicals in the pad may pollute the sea, and the seawater has a high probability of destroying the heating device. After the doll was heated with hot water, a comparison between the temperatures of body parts of a real person and of the doll was made with an infrared thermometer. The results show a temperature similarity between them, which proves the feasibility of using a clothed doll to simulate a human-like thermal appearance.

However, the heating strategy of the hot water poured on clothes has no ability to constantly warm the doll, which means the doll's surface temperature will decrease over time. Therefore, how the doll temperature changes along with time has to be investigated so that enough time (during which the doll remains at an optimal temperature for thermal imagery) is reserved for collecting the falling-into-water action—from the timing when the doll leaves the harbor edge to the timing when it hits the water. This investigation was done in three different scenarios: an indoor environment of 24 °C, and two outdoor environments of 17 °C and 0 °C, respectively. By measuring the doll's surface temperature along with time, we have found that the average decrease rates of the surface temperature are 0.03 °C/s, 0.06 °C/s, and 0.12 °C/s in the environments of 24 °C, 17 °C, and 0 °C, respectively. This has testified that all the three environments can reserve a time slot of at

least 3 minutes (before the doll temperature is decreased to the environment temperature) for performing a falling event.

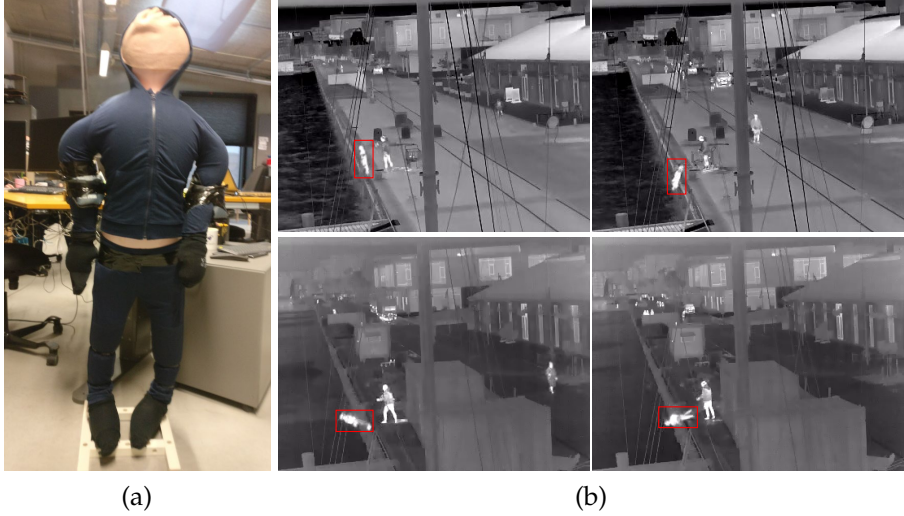


Fig. 3.5: The doll used for data collection and the sampled images from the gathered dataset. (a) The doll dressed in clothes. (b) Four sampled images showing the doll (in red bounding box) falling into water. Image source: [35].

After the preparation work mentioned above had been verified, we did the data collection in the harbor front by pushing the doll into water, throwing it into water, kicking it into water, letting it fall into water without exterior interference, etc. In this way, we have collected 22 thermal videos where 22 times of falling from the waterside were recorded. Each video has a frame rate of 25 fps, a resolution of 384×288 pixels, and a length varying from 5 seconds to 14 seconds. From these videos, we randomly sampled four images as shown in Fig. 3.5(b) to illustrate how the doll falls into water.

Not limited to making the doll's appearance in thermal imagery similar to humans, we have to further compare the motion features (from leaving the harbor to hitting the water) of the doll and a real person, to investigate whether this new falling-into-water dataset can be used in falling detection for drowning prevention in real life. We did this by comparing two aspects of the doll dataset and the person volunteers dataset in [16]: the number of frames and the optical flow in the period of leaving the waterfront edge to touching the water, which has shown the similarity in the falling motion of the doll and the volunteers.

Furthermore, due to the fact that a computer vision solution for drowning prevention usually inevitably requires human detection, the detection rate on the doll dataset by a human detector is thus a significant index to see whether this new dataset can contribute to preventing drowning accidents or not. By

4. Contributions

using YOLOv5 that is trained to detect humans, we find that it can detect the doll effectively with reliable confidence scores similar to what it does for detecting real persons.

In summary, our contributions regarding this falling-into-water dataset as part of paper H are:

- We have used a low-cost inflatable doll that is similar to a real person to simulate falling actions from a harbor edge to the water and thus collected a dataset that consists of 22 videos depicting 22 times of falling events. This data collection method not only avoids the problem of unconvincing simulation of generated artificial data by computer software but also reduces the potential injuries if human volunteers have to be involved in data collection.
- Both the appearance information and motion information of the doll dataset have been compared with those of real person volunteers. The similarity between them has proven that this falling-into-water dataset is suitable for drowning accident prevention in real life, from the perspective of the early detection (of a person suddenly disappearing from the harbor edge) instead of the late detection (of a person in water).
- For other researchers and engineers to use, this dataset has been made public at <https://www.kaggle.com/datasets/ivannikolov/thermal-mannequin-fall-image-dataset>.

For more detailed information of this dataset work, please refer to paper H in Part III.

Long-term Thermal Dataset as Part of Paper F [36]
















A generalization from a laboratory-tested algorithm to a real-life application always encounters a performance degradation, on account of the more complex and varied conditions in reality. Therefore, a computer vision solution that is targeted for the long-running surveillance system on *Safe Harbor* must be evaluated on a long-term thermal dataset that is similar to the scenes “seen” by the surveillance system as much as possible.

To this end, we have gathered such a long-term thermal dataset from the same surveillance system running in *Safe Harbor*, during the period from May 2020 to September 2020 and another period from January 2021 to April 2021. And then we have made the dataset public as part of paper F—*Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift* [36].

In general, the dataset is named as Long-term Thermal Drift (LTD) dataset as it includes many kinds of concept drift—the sudden, recurring, or gradual visual changes that happened during such a long period of eight months, comprehensively representing the real scenes of the certain harbor front in

Aalborg, Denmark. To give clearer illustration, Table 3.2 exhibits a few sampled images that demonstrate the concept drift from seasons, day and night shift, environments of different weather and human activities to show how the eight-month LTD dataset looks like.

Table 3.2: Sampled images of extreme concept drift contained in the LTD dataset. The right column “Environment” shows visualization changes based on human activities and weather situations which are many vehicles, many people, fog, rain, and snow from the first row to the last row, respectively. Adapted from paper F [36].

	Day	Night	Environment
Feb.			
Mar.			
Apr.			
Jun.			
Aug.			

In detail, the raw data of the gathered videos is too large for storage and

4. Contributions

publication. Thus the raw videos were split into shorter segments so that one 2-min clip was trimmed from each half-hour segment. This preprocessing not only decreases the resource consumption but also keeps the concept drift information that existing datasets do not have, resulting in the published version of the LTD dataset whose main information is listed in Table 3.3. Besides, to advance a further analysis of the LTD dataset, more metadata of it (listed in Table 3.4) is provided and made public.

Table 3.3: Main information of the LTD dataset. Adapted from the supplementary material of paper F.

Property	Value
Number of video clips	8940
Length of each clip	2 minutes
Length of all clips	298 hours
Time span	8 months
Season span	Spring, summer, winter
Format of each clip	MP4
Frame rate	25 fps
Spatial resolution	384×288

Table 3.4: Released metadata of the LTD dataset along with the video clips. Adapted from paper F and its supplementary material.

Metadata	Unit
Timestamp	year-month-day-hour-minute
Temperature	$^{\circ}\text{C}$
Humidity	%
Accumulated precipitation	kg/m^2
Dew point temperature	$^{\circ}\text{C}$
Wind direction	degrees
Wind speed	m/s
Mean sun radiation	W/m^2
Sunshine	minute

For its long-term property, the LTD dataset provides a new benchmark on which many computer vision models can be evaluated other than the usually used short-term benchmark datasets. Accordingly, to investigate how the performances of popular vision tasks will change on the dataset, six deep learning models corresponding to three fundamental tasks (two for autoencoders, two for anomaly detection, and the last two for human detection) have been tested on the LTD dataset with the following protocols:

- The two autoencoders are one 9-layer CNN-structured Classical Autoencoder (CAE) designed by us [37] and one Vector Quantized Variational Autoencoder (VQVAE2) [38]. The two models for anomaly detection are the reconstruction version and prediction version of the Memory-guided Normality for Anomaly Detection (MNAD) [39]. The two models for human detection are YOLOv5 [40] and Faster R-CNN [41].
- In this eight-month LTD dataset across seasons, the temperature change is one of the most significant variables and thus leads to a principal drift in thermal imagery. Therefore, for all the six models, the training sets come from the coldest February but are formed in three different variants—images sampled from the coldest day, the coldest week, and the whole month to study whether or not more varieties in training images help to improve the performance. The testing sets also have three variants—images sampled from January with a cold climate, April with a moderate climate, and August with a warm climate, respectively.
- Autoencoders and the models for anomaly detection are unsupervised studies that do not need manual annotations. Therefore, a total of 15000 frames for the training sets (with 5000 frames for each training variant) and 300 images for the testing sets (with each testing variant consisting of 100 images) are sampled from the LTD dataset. On the contrary, human detectors are supervised models that need annotated bounding boxes on humans. We, therefore, have annotated all the persons in 300 training frames (with each variant consisting of 100 images) and 300 testing frames (with each variant also consisting of 100 images) from the LTD dataset.
- For the two autoencoders and the two models for anomaly detection, the performance metric is the averaged Mean Square Error (MSE) over all the frames in the testing set. For the two human detectors, the performance metric is the detection rate mAP_{50} on the testing set.

Based on these protocols, we get the results in Table 3.5 and Table 3.6. From both tables, it is clear that when the testing images exhibit seasonal drifts from the training data, the performances of all the six models except the prediction version of MNAD decrease. This result meets the anticipation that a model trained on specific short-term data has difficulty in generalizing itself to a long-term application. The stable performance of MNAD Pred. is because this model needs a volume of consecutive frames as the input while the other five models only need one frame as the input; therefore, MNAD Pred. can get much richer and more robust input features than others.

Regarding the research question that whether or not more varieties in the training set can improve the performance, the two tables show that the train-

4. Contributions

ing images sampled from a week-level and a month-level indeed improve the performance in the MSE-measured models; but they do not improve the performance for YOLOv5 and Faster R-CNN. We attribute this difference to the reason that the annotated persons for human detectors may have minor variations no matter whether the frames are sampled from a shorter or longer period of time.

Table 3.5: Results of autoencoders and anomaly detection models measured by the averaged MSE over all the testing frames. Lower results indicate better performances. Table source: [36].

Methods	Train Feb.	Test		
		Jan.	Apr.	Aug.
CAE	Day 5k	0.0096	0.0202	0.0242
	Week 5k	0.0061	0.0167	0.0212
	Month 5k	0.0042	0.0109	0.0147
VQVAE2	Day 5k	0.0051	0.0072	0.0068
	Week 5k	0.0039	0.0066	0.0061
	Month 5k	0.0021	0.0039	0.0035
MNAD Recon.	Day 5k	0.0028	0.0057	0.0069
	Week 5k	0.0065	0.0066	0.0062
	Month 5k	0.0015	0.0041	0.0048
MNAD Pred.	Day 5k	0.0008	0.0007	0.0009
	Week 5k	0.0007	0.0006	0.0007
	Month 5k	0.0007	0.0006	0.0007

Table 3.6: Results of human detection models measured by the detection rate mAP_{50} over all the testing frames. Higher results indicate better performances. Table source: [36].

Methods	Train Feb.	Test		
		Jan.	Apr.	Aug.
YOLOv5	Day 100	0.8010	0.5390	0.5240
	Week 100	0.7940	0.4540	0.4860
	Month 100	0.7930	0.4860	0.4830
Faster R-CNN	Day 100	0.6760	0.3230	0.3370
	Week 100	0.6740	0.2790	0.3060
	Month 100	0.6400	0.2560	0.3180

Spontaneously, how to mitigate this effect of concept drift on computer vision tasks that decreases the performance is worthy of study. We correspondingly propose a solution as a baseline for it. To be specific, a novelty/outlier detector to find the time period during which the images are most different from others is applied, which can indicate the “time slot” of the prominent drift. By adding images from this slot to the training set, the performance on

the testing set can get an improvement. In our case, from March 5th 2021 (on which a large number of outliers were detected), we sampled and then added another 100 annotated images for human detectors and 5000 unannotated images for the other unsupervised models to redo the above evaluations. And the performances indeed get an improvement as expected.

Moreover, thanks to the metadata mentioned in Table 3.4, the exploration of which metadata factor accounts most for the performance drift is possible. We correspondingly did a correlation analysis and then found that the temperature and humidity have higher correlations to most of the model results than other factors, which encourages us to pay more attention to them when doing similar tasks in the future.

In summary, our contributions regarding this long-term thermal dataset as part of paper F are:

- We have collected a long-term thermal dataset from a real harbor surveillance system. This dataset consists of 8940 video clips from eight months during which seasonal, weather-caused, day and night shift-caused, and human activity-caused variations were recorded. The timestamp and climate conditions of each clip are also contained as metadata.
- This dataset provides a new benchmark to investigate how computer vision algorithms react to concept drift that exists in long-term periods, which helps to bridge the gap between short-term laboratory tests and long-running surveillance systems in the real world. Accordingly, we have tested six common computer vision models on the dataset and then proposed a baseline to mitigate the performance decrease due to concept drift.
- The accompanying metadata provides the material to explore the relationship between the performance of computer vision tasks and various environmental factors, based on which people can pay more attention to some specific environmental aspects for a better and stable result. Accordingly, we have found that the temperature and humidity have a great influence on performances in our case.
- For other researchers and engineers, this long-term thermal dataset, together with the training and testing sets for the six computer vision models, is made public at <https://www.kaggle.com/datasets/ivannikolov/longterm-thermal-drift-dataset>. The accompanying codes are also made public at <https://github.com/IvanNik17/Seasonal-Changes-in-Thermal-Surveillance-Imaging>.

For more detailed information of this dataset work, please refer to paper F in Part III.

4.2 Drowning Accident Prevention: Warning in Advance

Compared to conditions of swimming pools, a harbor front has a much larger area, and the lifeguards are on standby in a relatively farther away place. This determines the importance of a warning in advance to indicate a potentially drowning accident. With it, authorities or professionals can pay beforehand attention to the person in danger and get prior knowledge of his or her location. Thus, such an idea of warning in advance will significantly reduce the preparation time for rescue and then reduce the casualties, especially under the anticipation that a successful rescue has to be performed within a few short minutes.

Therefore, different from existing works on drowning detection in the water, we prevent drowning accidents by sending out alarms when detecting a person very near to the waterside. This idea is viable under an observational finding that people tend to walk away from the harbor edge when there is no accompanies during the night (Fig. 3.6); therefore, there will be a very limited number of false alarms. On the other hand, a situation where a person is within a waterside region without witnesses nearby is most dangerous and may cause drowning accidents, for which an alarm in advance is in dire need.

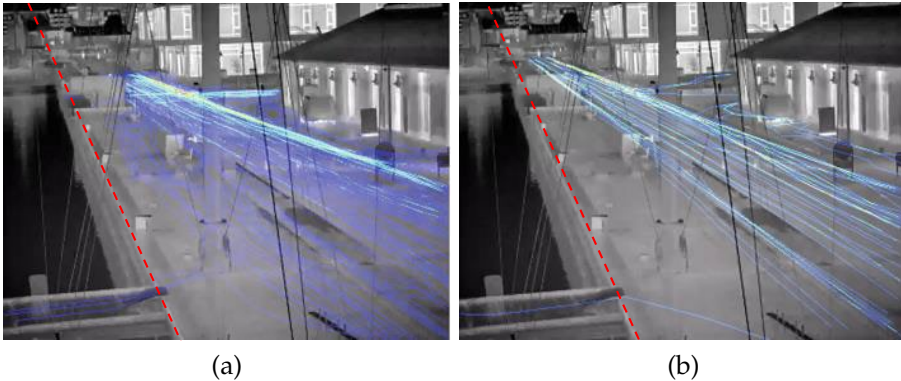


Fig. 3.6: Trajectories of 237 persons in the daytime (a) and of 42 persons at night (b). The more pedestrian traffic, the larger density in trajectories with more intense colors. The red dashed line locates the harbor edge and the left of it is the water area. Image source: [16].

To meet such a need in computer vision, paper E, *Supervised versus Self-supervised Assistant for Surveillance of Harbor Fronts* [37], has proposed two alternative solutions:

- Supervised human detection: A human detector can give the information of a person's location and the distance from the waterside, based on which an alarm will be raised or not.

- Self-supervised anomaly detection: A predefined region very near to the water is the input of an autoencoder for anomaly detection. The autoencoder is first trained with only normal data where there is no person in it. And then, during the application phase, the autoencoder can help to detect anomalous inputs where there is human activity near the waterside by using the reconstruction error as an indicator.

To apply the supervised human detection approach, we decided to use YOLOv5 [40] object detector for its powerful feature reuse ability by means of PANet [42] and its multiple-scale outputs that fit the multiple sizes of people from far and near. When a person is detected by YOLOv5, the next step is to decide if he or she is in the alarm region based on the detected location of him or her. Accordingly, the alarm region of the specific harbor front in *Safe Harbor* has to be defined in advance as the left area of the red line shown in Fig. 3.7. This definition is empirical on the basis of the fieldwork in the harbor front and the long-time observations of the captured scenes.

In mathematics, the image captured by the surveillance camera constructs an image coordinate system, also shown in Fig. 3.7. The origin of the system is the left bottom, and the image canvas is a 2D plane from $(0, 0)$ to $(383, 0)$ on x -axis and from $(0, 0)$ to $(0, 287)$ on y -axis, considering that the thermal camera sensor has a spatial resolution of 384×288 pixels. In this coordinate system, the red boundary line is defined in equation 3.1. For a detected person represented by a bounding box from YOLOv5, if any pixel with coordinates (x_p, y_p) in the bounding box satisfies the formula 3.2, the person is considered in the alarm region that may result in a falling into water accident.



Fig. 3.7: The red line indicates the alarm boundary and the left of it is defined as the alarm region. Image source: [37].

4. Contributions

$$1.53x + y - 283 = 0 \quad (3.1)$$

$$1.53x_p + y_p - 283 < 0 \quad (3.2)$$

To apply the self-supervised anomaly detection approach, we have designed and implemented a 9-layer CNN-structured autoencoder in which a 5-layer encoder and a 5-layer decoder share a layer of bottleneck. To comply with the alarm region defined for the supervised human detection approach, the input of the autoencoder is also a specifically defined image region shown in Fig. 3.8, that is, the region from the water edge to the alarm line. To make this region in a rectangular shape of size 64×192 that can be fed to the autoencoder, a further transformation is done for this region, which results in the red box area on the right side of the figure.



Fig. 3.8: The region from the water edge to the alarm line is defined as the input for the autoencoder to detect anomalies. To make the region rectangular, a transformation is then required. Image source: [37].

The autoencoder is first trained with those 64×192 rectangles where there is no human. While in the testing phase, any 64×192 rectangle with or without persons in it will be fed to the well-trained autoencoder. As the autoencoder has never seen a person in its training phase, it cannot reconstruct any person patterns. Therefore, for an input with humans in it, the reconstruction error (in the form of MSE) between the input and the reconstructed output will be higher than that of an input without a human in it.

In a real application, defining a threshold in advance is required for the aim that an input with its MSE above the threshold is detected as an anomaly. By applying the autoencoder on a small annotated dataset from which each image is labelled with normal or abnormal, a threshold that achieves a good balance between the detection precision and the recall rate can be acquired.

And then, for the testing data, if an input has its MSE larger than the threshold, it will be detected as an anomaly referring to human activities in the alarm region.

To evaluate these two schemes, corresponding training sets and testing sets have been prepared:

- 2358 thermal images were sampled from the videos captured by the harbor surveillance system from February 3th 2020 to March 3th 2020. All the persons in these images were then annotated with a bounding box around for YOLOv5. As a result, the annotated dataset was further split into a training set of 1715 images, a validation set of 143 images, and a testing set of 500 images.
- The same 2358 images were used to train and evaluate the designed autoencoder. The same 1715 images for training YOLOv5 were labelled into 87 abnormal images with persons in the alarm region and other 1628 normal images without persons in the alarm region. On this 1715 labelled dataset, a threshold is obtained.
- For the final goal of detecting anomalies, the testing set of 500 images was labelled into 91 abnormal images and 409 normal images for the evaluation purpose.

Correspondingly, on the above-prepared datasets, YOLOv5 achieves a human detection rate of 97.70% measured by mAP_{50} . Based on the human detection result and the formula 3.2, 85 images are detected as abnormal out of the whole 91 testing images, and no false alarm is raised, equaling a precision rate of 100% and a recall rate of 93.41%. On the other hand, based on the acquired threshold, the autoencoder (trained with 1628 normal images) detects 103 images as abnormal, but there are 21 false alarms, equaling a precision rate of 79.61% and a recall rate of 90.11%.

From these results, an initial thought is that the human detection approach is better than the anomaly detection approach. A further analysis of the failed cases of the autoencoder has been done, which points out two dominant problems: (1) false alarms are caused by non-living objects with high temperatures, as shown in Fig. 3.9(a); (2) missing detections will happen when a person is going from the safe region to the alarm region, because only a very tiny part of the body is captured in the 64×192 sized input, as shown in Fig. 3.9(b). After a manual double-check to filter out the cases similar to Fig. 3.9(b) (people are just entering the alarm region from the safe region), the performance of the autoencoder-based anomaly detection improves a lot from 0.929 to 0.995 measured by the area under the precision-recall curve (AUC).

In summary, our contributions regarding this warning in advance idea as the key innovation of paper E are:

4. Contributions

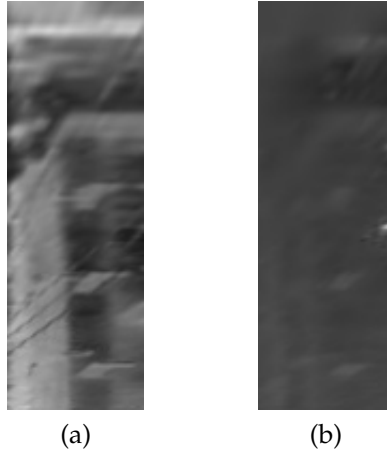


Fig. 3.9: A false alarm (a) caused by the harbor front metals and concretes that have a high temperature after a long day of direct solar radiation. A missing detection (b) that a person is going from the safe region to the alarm region indicated by the very few brighter pixels near the right boundary. Image source: [37].

- Different from existing works on preventing drowning accidents that detect persons in the water, we have proposed a “one-step-earlier” strategy that sends out alarms when detecting humans very near to the waterside. This makes lifeguards or other professionals pay beforehand attention to the situation in advance to prevent a drowning event from happening or provide faster rescues.
- Accordingly, we have proposed one human detection scheme and one anomaly detection scheme to detect the potentially dangerous scenes, and both of them are proved effective.
- For other researchers and engineers to refer to and use, the codes and the labelled training and testing sets are made public at <https://github.com/JinsongCV/Supervised-Versus-Self-supervised-Assistant-for-Surveillance-of-Harbor-Fronts>.

For more detailed information of this warning in advance work, please refer to paper E in Part III.

4.3 Anomaly Detection: on Long-term Data with Concept Drift

Anomaly detection, as mentioned before, is often realized by the strategy that trains an autoencoder with only normal data, which makes it unfamiliar with anomalous data. This “unfamiliarity” is presented by the reconstruction error

between the input and its output. The larger error, the larger probability the input is an anomaly.

However, this strategy only works provided that the definition of normal data is fixed. A long-running surveillance system is hardly to guarantee this prerequisite due to concept drift across time. Therefore, the evolution from an existing anomaly detector evaluated on short-term datasets to a new version that can be applied to long-term systems is in dire need. The expectation is that the improved anomaly detector must have the ability to distinguish real anomalies from irrelevant changes resulting from concept drift.

For this aim, paper G, *Detecting Anomalies Reliably in Long-term Surveillance Systems* [43], has proposed a weighted reconstruction error strategy to get robust results on long-term data. Specifically, the development and evaluation are done on the collected LTD dataset [36] as it is a good representation of a long-running surveillance system in real life—considering the diverse concept drift phenomena in the LTD dataset as exemplified in Fig. 3.10.

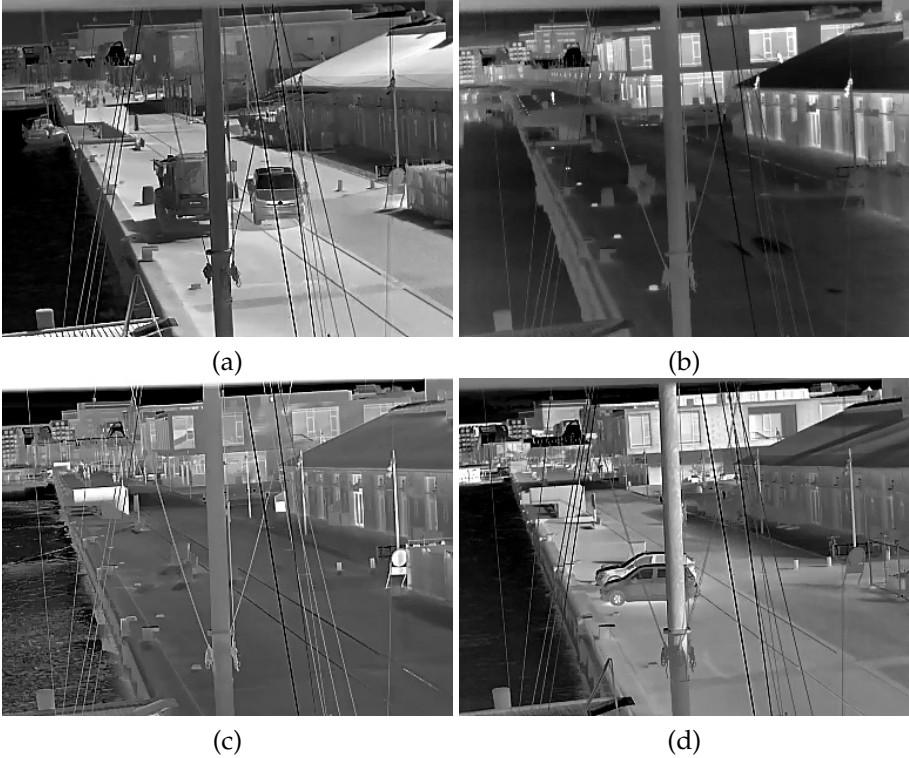


Fig. 3.10: All the images are normal data without dangerous/harmful human activities but show significant visual changes: (a) August, (b) January, (c) February, (d) April. Image source: [43].

In detail, the diagram of the proposed anomaly detection scheme is in

Fig. 3.11 where the reconstruction error between the input and the output is calculated by a pixel-by-pixel difference in the form of MSE. The red pipeline refers to the traditional scheme in which the weight of each pixel is the same. Therefore, all the pixels (no matter whether they are from the background or the foreground) will have an equal contribution in calculating the MSE value, which inevitably considers the influence of environmental drifts (like the contrast change and the sea ripples in Fig. 3.10) into the MSE as well. In this way, the long-term input with concept drift will make the MSE curve fluctuate across time, which will cause the MSE values of normal data to be even higher than that of anomalies and thus induce missing detections.

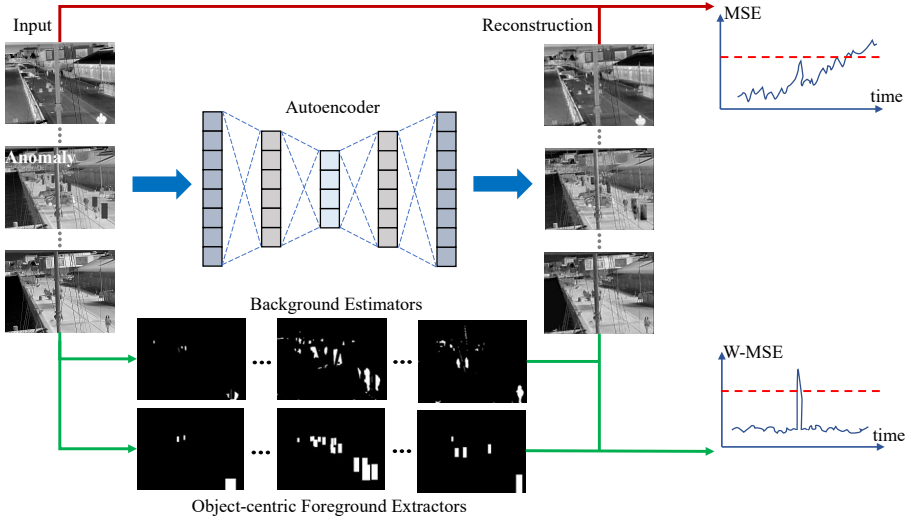


Fig. 3.11: The diagram of the proposed anomaly detection scheme with weighted reconstruction error as its main idea. The red flow is the traditional pipeline. The green flow is the proposed pipeline. The red dashed line refers to the predefined threshold to detect an anomaly, and an input with its MSE above it will be detected. Image source: [43].

On the contrary, the green pipeline in Fig. 3.11 refers to the proposed scheme that uses the weighted reconstruction error strategy. This scheme additionally introduces a foreground/background segmentation module, and thus different weights can be assigned to the foreground pixels and background pixels in calculating the reconstruction error MSE. By giving foreground pixels larger weights, the MSE value can pay more attention to the human activity region where anomalies occur in default and thus, at the same time, mitigate the influence of the environmental changes in the background region. In this way, the weighted MSE curve will not fluctuate as much as the MSE curve of the conventional scheme for a long-term input, making the detection of anomalies much more accurate by detecting peaks that are above the threshold. To be noted is that the diagram is specifically depicting the

inference phase.

To evaluate the proposed weighted reconstruction error strategy for anomaly detection, paper G implemented the corresponding autoencoders and the foreground/background segmentation modules and prepared the required dataset:

- Three autoencoders are applied: one Vector Quantized Variational Autoencoder (VQVAE2) [38], one autoencoder using Memory-guided Normality for Anomaly Detection (MNAD Recon.) [39], and one CNN-structured Classical Autoencoder (CAE) composed of an 11-layer encoder and an 11-layer decoder designed by us.
- Five kinds of foreground/background segmentation methods are applied: mixture of Gaussians (MOG2) for background estimation [44], mixture of Gaussians using K-nearest neighbours (KNN) for background estimation [45], image difference with arithmetic mean (ID_a) for background estimation, image difference with Gaussian mean (ID_g) for background estimation, and object-centric foreground estimation by use of the YOLOv5-based human detection [40].
- The segmentation results of a thermal image are shown in Fig. 3.12 in which (b)-(e) demonstrate foreground regions with grayscale values near 255 and background regions with values near 0. In contrast, YOLOv5-based segmentation in Fig. 3.12(f) demonstrates each foreground region as a bounding box with a constant grayscale value (which equals the multiplication of the corresponding person's detection confidence score and 255), while the background regions are with grayscale value 0.
- For any input I whose segmentation map is M , the weight of each pixel (I_x, I_y) for calculating the reconstruction error is the quotient of a division. The dividend of the division is the grayscale value of the corresponding pixel (M_x, M_y) from M , and the divisor is value 255. This weight assignment method is the simplest way to increase the foreground region weight and weaken the background region weight in the MSE calculation. It is worth mentioning that the proposed weighted reconstruction error strategy is not limited to this weight assignment scheme; more complex schemes like a combination of multiple segmentation results are also possible, which have been given in paper G.
- The three autoencoders (VQVAE2, MNAD, and CAE) are trained with images from February 2021 as part of the LTD dataset. The testing sets have two versions: 300Ver of 300 images sparsely sampled from April 2021, January 2021, and August 2020, and 3515Ver of 3515 images densely sampled from the same months for an extended evaluation.

In 300Ver, there are 78 anomaly images; and in 3515Ver, there are 60 anomaly images.

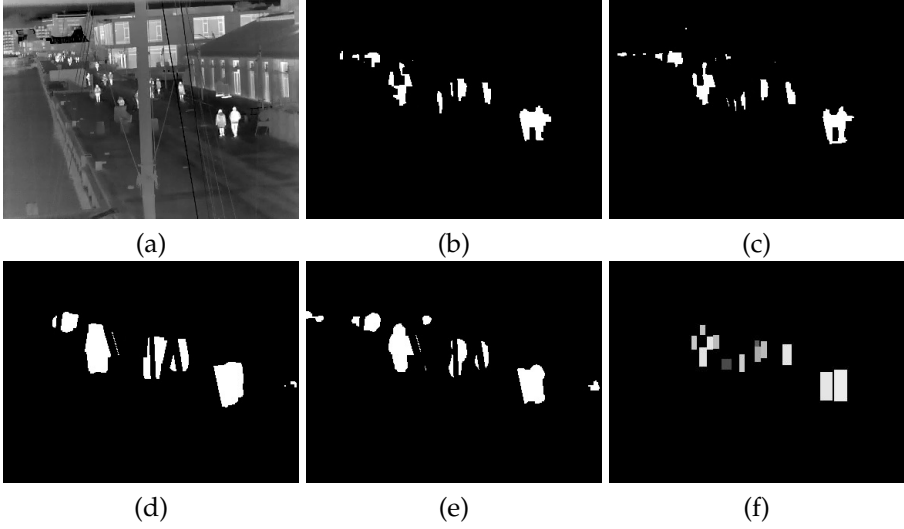


Fig. 3.12: One thermal input and the foreground/background segmentation results from the implemented five methods. (a) Input. (b) MOG2. (c) KNN. (d) ID_a . (e) ID_g . (f) YOLOv5. Image source: [43].

To show the evaluation result, we draw how the MSE curve changes when applying VQVAE2 to the 300Ver testing set in Fig. 3.13. In it, a new term GT refers to ground truth, and the GT curve means the version that the segmentation is based on the manually annotated human bounding boxes. Obviously, all the MSE curves in Fig. 3.13(b) that either use the conventional MSE calculation or pay more attention to the background region have a highly-similar trend, demonstrating that the conventional MSE calculation cannot accurately detect anomalies which usually happen in the foreground area. In contrast, the six MSE curves in Fig. 3.13(a) that pay more attention to the foreground region have totally different trends from the curves in Fig. 3.13(b), demonstrating that the weighted MSE idea can avoid the influence of environmental changes that significantly degrade the robustness of the conventional MSE calculation scheme. Therefore, the proposed pipeline has the ability to do anomaly detection reliably on long-term datasets with concept drift.

To further verify this ability, the detection rate of VQVAE2 on 300Ver has been calculated by finding how many anomalies are detected from the images of the largest 10% MSE values. And hence we get the anomaly detection rates of the conventional method, MOG2-based weighted method, KNN-based weighted method, ID_a -based weighted method, ID_g -based weighted

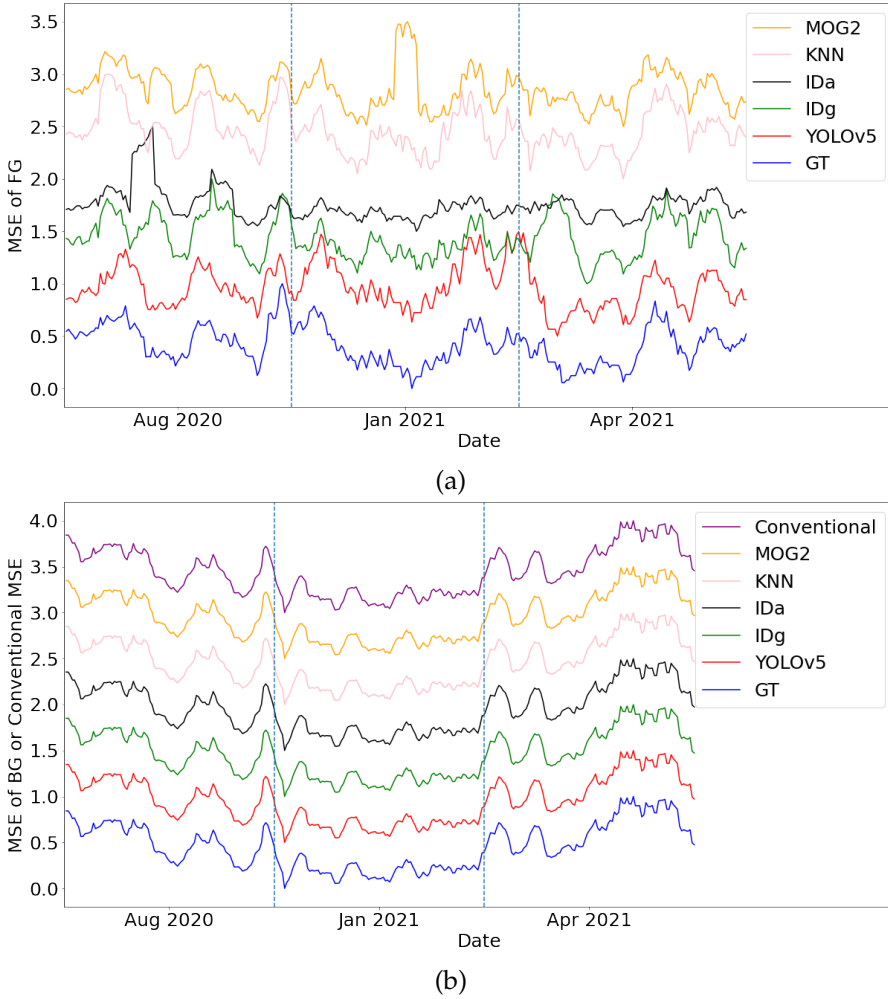


Fig. 3.13: (a) Weighted MSE curves that pay more attention to the foreground (FG). (b) Weighted MSE curves that pay more attention to the background (BG) and a conventional MSE curve. These MSE curves across time are drawn after a smoothing step of a mean filter whose kernel size is 10, a step of normalization to $[0, 1]$, and a step of translation to avoid the overlapping with other curves. We use some vertical dashed azure lines to separate months of August, January, and April. Image source: [43].

method, and YOLOv5-based weighted method as 24.36%, 69.23%, 71.79%, 65.38%, 65.38%, and 78.21%, respectively. Similar results are also observed when applying MNAD and CAE on 300Ver and from the extended experiments on 3515Ver. All of these results prove the improvement of the proposed weighted-MSE strategy in anomaly detection.

In summary, our contributions regarding this weighted reconstruction er-

rior idea from paper G are:

- To solve the problem that existing anomaly detection evaluated on short-term datasets may be problematic for long-term applications with concept drift, we have proposed a “pay attention to what is crucial” strategy.
- The specific idea is that the background pixels and foreground pixels have different weights in calculating the reconstruction error between an input and its output of an autoencoder, which can assign an extra focus on human activities that an anomaly in default relates to. In this way, the drawback that environmental drifts are also counted as part of the reconstruction error in conventional anomaly detection pipelines is overcome to some extent, and thus more reliable anomaly detection results are possible.
- We have used three kinds of autoencoders and two datasets spanning three months to evaluate our idea and proven that the proposed method is a more robust solution for anomaly detection in long-term applications that run for months and years.
- For other researchers and engineers, the codes and the accompanying datasets are made public at <https://github.com/JinsongCV/Weighted-MSE>.

For more detailed information of this anomaly detection work developed for long-term datasets, please refer to paper G in Part III.

Sub-conclusion

In the project of *Safe Harbor*, we not only pay attention to the most dangerous accident on harbor fronts—drowning (papers E and H) but also keep an eye on other incidents that may induce harmful or potentially dangerous consequences (papers F and G).

Specifically, as every second counts in rescuing drowning persons, an alert system in advance instead of finding persons in the water is more important. For this purpose, detecting persons very near the waterfront (paper E) and detecting the falling event from a harbor edge (paper H) are useful solutions. Furthermore, rarely-happening incidents (fighting, vehicle collision, anomalous crowds, etc.) in the harbor front need immediate control too. For this purpose, a robust anomaly detector (paper G) that has to work in an environment with concept drift (paper F) is developed. We expect these four works together to provide a solution to improve the safety level in a waterfront area.

References

- [1] Vassilios Tsakanikas and Tasos Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736–753, 2018.
- [2] Guruh Fajar Shidik, Edi Noersasongko, Adhitya Nugraha, Pulung Nurtantio Andono, Jumanto Jumanto, and Edi Jaya Kusuma, "A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and datasets," *IEEE Access*, vol. 7, pp. 170457–170473, 2019.
- [3] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, pp. 103116, 2021.
- [4] Romas Vijeikis, Vidas Raudonis, and Gintaras Dervinis, "Towards automated surveillance: A review of intelligent video surveillance," *Intelligent Computing*, pp. 784–803, 2021.
- [5] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.
- [6] Lingmei Ren and Yanjun Peng, "Research of fall detection and fall prevention technologies: A systematic review," *IEEE Access*, vol. 7, pp. 77702–77722, 2019.
- [7] Fengju Bu and Mohammad Samadi Gharajeh, "Intelligent and vision-based fire detection systems: A survey," *Image and vision computing*, vol. 91, pp. 103803, 2019.
- [8] Aalborg University, "Baggrund for projektet," <https://tryghavn.create.aau.dk/index.php/baggrund-for-projektet/>, last accessed: March, 2022.
- [9] World Health Organization et al., "Global report on drowning: preventing a leading killer," 2014.
- [10] Wenmiao Lu and Yap-Peng Tan, "A vision-based approach to early detection of drowning incidents in swimming pools," *IEEE transactions on circuits and systems for video technology*, vol. 14, no. 2, pp. 159–178, 2004.
- [11] Nasrin Salehi, Maryam Keyvanara, and Seyed Amirhassan Monadjemmi, "An automatic video-based drowning detection system for swimming pools using active contours," *IJ Image, Graphics and Signal Processing (IJIGSP)*, pp. 1–8, 2016.

- [12] Abdel Ilah N Alshbatat, Shamma Alhameli, Shamsa Almazrouei, Salama Alhameli, and Wadhha Almarar, "Automated vision-based surveillance system to detect drowning incidents in swimming pools," in *2020 Advances in Science and Engineering Technology International Conferences (ASET)*. IEEE, 2020, pp. 1–5.
- [13] P Pavithra, S Nandini, A Nanthana, Noor Tabreen Aslam, and Praveen Kumar, "Video based drowning detection system," in *2021 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*. IEEE, 2021, pp. 203–206.
- [14] Yi-Tung Chan, Tai-Wei Hou, Yu-Lun Huang, Wen-Hsin Lan, Pin-Chia Wang, and Cih-Ting Lai, "Implementation of deep-learning-based edge computing for preventing drowning," *computing*, vol. 8, pp. 13.
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [16] Soren Bonderup, Jonas Olsson, Morten Bonderup, and Thomas B Moeslund, "Preventing drowning accidents using thermal cameras," in *International Symposium on Visual Computing*. Springer, 2016, pp. 111–122.
- [17] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.
- [18] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.
- [19] Jürgen Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
- [20] Yong Shean Chong and Yong Haur Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International symposium on neural networks*. Springer, 2017, pp. 189–196.
- [21] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.
- [22] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, "Future frame prediction for anomaly detection—a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.

- [23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [24] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.
- [25] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.
- [26] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.
- [27] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7842–7851.
- [28] Mariana-Iuliana Georgescu and Marius Popescu Mubarak Shah Radu Tudor Ionescu, Fahad Shahbaz Khan, "A scene-agnostic framework with adversarial training for abnormal event detection in video," *arXiv*, 2020.
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multi-box detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [30] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [31] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1975–1981.
- [32] Ramin Mehran, Alexis Oyama, and Mubarak Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.

- [33] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, 2008.
- [34] HIKVISION, "Ds-2td2235d-25/50 thermal network bullet camera," <https://us.hikvision.com/en/products/more-products/discontinued-products/thermal-camera/thermal-network-bullet-camera-ds>, last accessed: March, 2022.
- [35] Ivan Nikolov, Jinsong Liu, and Thomas Moeslund, "Imitating emergencies: Generating thermal surveillance fall data using low-cost human-like dolls," *Sensors*, vol. 22, no. 3, pp. 825, 2022.
- [36] Ivan Adriyanov Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Velling Dueholm, Anders Skaarup Johansen, Kamal Nasrollahi, and Thomas B Moeslund, "Seasons in drift: A long-term thermal imaging dataset for studying concept drift," in *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.
- [37] Jinsong Liu, Mark P Philipsen, and Thomas B Moeslund, "Supervised versus self-supervised assistant for surveillance of harbor fronts," in *VISIGRAPP (5: VISAPP)*, 2021, pp. 610–617.
- [38] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.
- [39] Hyunjong Park, Jongyoun Noh, and Bumsu Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.
- [40] Glenn Jocher et. al., "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," <https://doi.org/10.5281/zenodo.6222936>, Feb. 2022.
- [41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.
- [42] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

References

- [43] Jinsong Liu, Ivan Nikolov, Mark Philipsen, and Thomas Moeslund, "Detecting anomalies reliably in long-term surveillance systems," in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP. INSTICC*, 2022, pp. 999–1009.
- [44] Zoran Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, 2004*, vol. 2, pp. 28–31.
- [45] Zoran Zivkovic and Ferdinand Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, no. 7, pp. 773–780, 2006.

Chapter 4

Summary

The PhD thesis studies how to apply the latest computer vision ideas and techniques to two projects *Thermal Adaptive Architecture* and *Safe Harbor*. Though the projects are different in many aspects, both are interested in analyzing people in thermal imagery. In this overview, the background, introduction, and related work of each project have been narrated. Besides, our contributions to each project are also described by introducing the key contents of our publications appended in the following parts. Together they have answered the questions we raised in Chapter 1.

Thermal Adaptive Architecture

Q: What factors influence a person's thermal sensation?

A: According to ISO standards, four environmental factors of air temperature (T_a), mean radiant temperature (\bar{t}_r), air velocity (V_a), and relative humidity (RH); two personal factors of clothing insulation rate (I_{cl}) and metabolic rate (M). Besides, the gender difference also has an influence on individual thermal sensation.

Q: How to acquire these factors from computer vision solutions?

A: The environmental factors are measured easily by sensors. The acquisition of the personal factors of I_{cl} and M requires three modules: (1) a human detection and tracking module by YOLOv5 and DeepSort to track each individual, (2) a key body parts detection module by OpenPose for the measurement of clothes temperature and skin temperature based on which to calculate I_{cl} , (3) an optical flow calculation module based on which to predict the person's activity level that indicates the M value. The recognition of gender can be achieved by a CNN-based classifier.

Q: Compared with existing manual work, are the acquired factors from computer vision solutions sufficient and accurate?

A: All the computer vision modules mentioned above have been evaluated on corresponding data and thus been proven that they together provide an automatic solution with both convenience and accuracy than manual work. Detailed information of the performances and analyses are in the papers and the technical report in Part II.

Safe Harbor

Q: What anomalies should be considered and detected?

A: A specific anomaly of drowning and other general anomalies with dangerous or harmful consequences (like traffic accidents, crowds during an epidemic, fights, etc.) that need extra attention and further controls or rescues from professionals.

Q: How to detect the considered anomalies from computer vision solutions?

A: A human detector can give the information of a person's location and the distance to the waterside, based on which we can predict if the person is in an alarm region that may lead to the anomaly of drowning. A general anomaly detector structured from an autoencoder can detect an anomaly by marking it as "unfamiliarity" which is represented by a large reconstruction error between the autoencoder's input and output.

Q: Is the detection method fast and efficient for a timely control or rescue?

A: To give a fast and efficient response to anomalies that require extra human attention, we not only consider an early-alarm strategy to reduce the preparation time for a rescue but also develop anomaly detection algorithms on a long-term thermal drift dataset. This new dataset has very similar properties to a running surveillance system in real life, unlike other short-term datasets that are far away from the real conditions. We believe both aspects are beneficial for a faster and stronger method.

As a whole, this complementarity of the indoor research *Thermal Adaptive Architecture* and the outdoor research *Safe Harbor* makes the PhD thesis more comprehensive and decidedly claims that

With the automatic analysis of people in thermal imagery, computer vision techniques can ease manpower for a more comfortable and safer life.

Future Work

Though we have achieved the above-summarized results, there are many other works in the future worthy of study, which are introduced as follows.

Thermal Adaptive Architecture

From this PhD work, an automatic and dynamic estimation of individual clothing insulation rate and metabolic rate is developed, which makes it possible to assess each person's thermal feeling in real time in an indoor microclimate. Therefore, in the future, we plan to conduct closer cooperation with researchers and engineers from the architecture design field so that our work can be applied to control the HVAC systems, intelligent curtains, or other facilities. In this way, the direct regulation of the temperature in separate local areas is possible, which can meet the requirements of different subjective thermal sensations. Besides, regarding the ongoing work—gender classification in thermal imagery, we plan to extend existing experiments and analyses to more thermal datasets to further verify what we have discovered. For this plan, a modality conversion of transforming existing RGB benchmark datasets to the thermal modality using generative adversarial networks will be studied, considering that the amount and diversity of visible datasets are much larger and richer than thermal modality databases.

Safe Harbor

From this PhD work, by developing algorithms on the long-term thermal drift dataset, raising early warnings for preventing drowning accidents and detecting other emergencies or potentially dangerous incidents are possible. In the future, we plan to extend these researches on the dataset to the real running surveillance system to detect falling events into water, traffic accidents, anomalous crowds, etc., so as to improve the safety level in the harbor front in real life. Besides, regarding the concept drift phenomenon which has a negative impact on almost all computer vision tasks, we plan to propose more solutions to mitigate this influence on object detection and anomaly detection in addition to those that have been studied in papers F and G.

Chapter 4. Summary

Part II

Thermal Adaptive Architecture

Paper A

Vision-based Individual Factors Acquisition for Thermal Comfort Assessment in a Built Environment

Jinsong Liu, Isak Worre Foged, and Thomas B. Moeslund

The paper has been published in the
*Proceedings of the 15th IEEE International Conference on Automatic Face and
Gesture Recognition (FG), 2020*

© 2020 IEEE

Reprinted, with permission, from Jinsong Liu, Isak Worre Føged, and Thomas B. Moeslund. “Vision-based individual factors acquisition for thermal comfort assessment in a built environment.” 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020, pp. 662-666.

The layout has been revised.

Abstract

To maintain satisfactory chamber thermal environments for occupants, heating, ventilation and air conditioning (HVAC) systems have to work frequently. However, the room conditions especially the temperatures are usually set empirically which fail to consider occupants' real needs, not to mention personalized thermal comfort, therefore, the HVAC systems are underutilized and unavoidably induce energy waste. To solve this problem, a vision-based method to acquire multiple individual factors that are critical for assessing personalized thermal sensation is proposed. Specifically, with the indoor videos captured by a thermal camera as inputs, a convolutional neural network (CNN) is implemented to recognize an occupant's clothes and action type simultaneously. With a dataset of 20 persons, the experimental results show an average classification rate of 95.14% on 4 dataset partitions for a 15-category scenario, which prove the effectiveness of the proposed method.

1 Introduction

People spend most time indoors, and heating, ventilation and air conditioning (HVAC) systems, therefore, have to operate frequently to maintain satisfactory indoor thermal conditions for occupants thus improving their life quality. Though current HVAC systems do have controllers, the environment is usually set with fixed empirical temperatures like 26°C in summer and 20°C in winter, which ignores the dynamic thermal needs of individuals and also leads to overcooling or overheating with great energy waste.

Researchers have tried to solve this problem in various ways. In the 1960s, Fanger [1] first defined 7-scale thermal sensations (cold, cool, slightly cool, neutral, slightly warm, warm and hot) and gave a predicted mean vote model that estimated a group of persons' average thermal sensation in a certain indoor condition. This model is determined by six factors: four environmental values (air temperature, mean radiation temperature, relative humidity and air velocity) and two personal values (clothing rate (CLO) and metabolic rate (MET)).

However, as Fanger's model is based on college-aged students in a constant indoor environment of moderate thermal climate zones, it failed to generalize to other situations, and thus several improvements have emerged. For instance, the questionnaire-based methods record age, gender, CLO, MET and thermal feelings manually in various indoor environments, making the modified models more applicable to different situations; the measurement-based methods rely on devices to measure an occupant's thermal-related attributes, which mainly pay attention to human body temperature acquisition for its simplicity.

Whatever the approaches are, they all accept the significance of the

Table A.1: CLO values of some clothes [2].

Clothes Type	T-shirts	Short-sleeve shirt	Long-sleeve shirt
CLO	0.08	0.19	0.25

Table A.2: MET values of some actions [1].

Action Type	Sitting	Standing	Normal walking
MET	1.0	1.4	2.6

above-mentioned six factors in assessing an occupant's thermal comfort level. Among them, the four environmental factors can be measured by the thermometer, hygrometer and anemograph. The remaining personal factors are usually ignored due to the difficulty of acquisition. To overcome this problem, we propose a vision-based approach to simultaneously obtain an occupant's clothes type and action type which correspond to specific CLO and MET values, like what Table A.1 and Table A.2 list. The concrete contributions are:

- A contactless and privacy-preserving method is proposed to acquire multiple personal factors with only one thermal camera as the input source.
- A convolutional neural network (CNN) considering both spatial and temporal information to classify an occupant's clothes and action type is implemented, achieving an average classification rate of 95.14% for a 15-category (5 clothes types and 3 action types) thermal dataset.
- Comprehensive experiments with different scenarios are given to advise good practices for such a task.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the proposed method. Section 4 presents the experiments. Section 5 concludes this paper.

2 Related Work

For a more accurate thermal comfort assessment, several works have been published. Questionnaire-based methods [3–5], are a simple and direct way to know a person's thermal state. However, this method is impractical in modern applications as it interferes with subjects. Therefore, the related works we focus on are measurement-based methods, further subdivided into

3. Proposed Method

contact, semi-contact and contactless methods according to the employed device types.

Contact methods refer to using devices attached to the skin to detect human body physiological parameters. Sim [6], Wu [7] and Chaudhuri [8] utilized the thermistor (Lattron LNJT103F), the surface thermometer (Testo 905-T2) and the skin thermometer (Exacon D-S18JK) respectively to acquire skin temperatures of different body locations and then explored the relationship between an occupant's thermal feelings and the detected temperatures.

Semi-contact methods refer to using wearable devices to detect human body physiological parameters. Ghahramani [9, 10] used an infrared sensor installed on the eyeglass frame to measure the change of skin temperatures on a human face under different conditions. Na [11] employed a smart bracelet (Fitbit Charge 2) to monitor a person's heart rate variation continuously to discover the connection between activities and heart rates thus predicting the MET.

Even though the above two categories contribute to the development of thermal comfort understanding, their practical applications are immensely limited due to the inconvenience caused by adhered or wearable devices. Contactless methods, therefore, stand out. Cheng [12] proposed a method that extracts the saturation channel of images of the hand captured by a normal RGB camera; as this channel information is closely relevant to the skin temperature, it can be used to predict thermal sensations directly without extra temperature extraction. Li [13] presented a thermal camera-based framework that measures temperatures of different facial parts and found that ears, nose and cheeks are most indicative of one's thermal comfort levels. Lu [14] provided an RGB camera-based system for activity level classification realized by finding pixels with large deviations which indicate people moving to control the HVAC switch.

In summary, the existing methods usually focus on only one aspect, either human body temperatures or activity situations, not to mention an integrated system to acquire multiple personal factors. In other words, lots of useful information is missing. To fill the gap, we provide a method that simultaneously acquires an occupant's clothes and action type for further CLO and MET estimation, thus benefiting individual thermal comfort assessment. To the best of our knowledge, the proposed approach is the first comprehensive work in this research field.

3 Proposed Method

Clothes type and action type can be recognized via two separate phases, but a single pipeline predicting both tasks is much more efficient and practically useful. The overview of our proposed method can be seen in Fig. A.1.

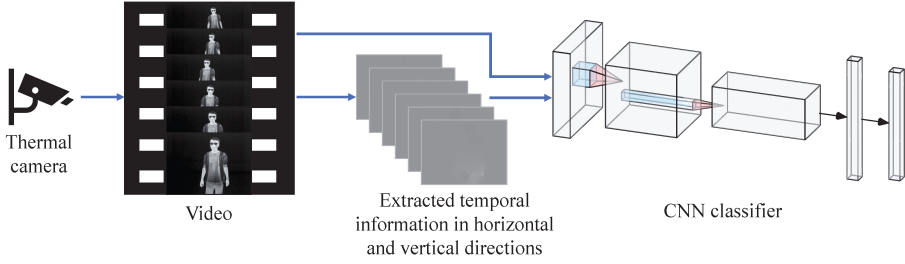


Fig. A.1: Overview of the proposed method.

For action recognition, the two-stream convolutional networks [15] as well as its improvements like the hidden two-stream convolutional networks [16] and the temporal segment networks (TSNs) [17] lay a solid direction for this task. All of them rely on two streams, one is the spatial flow with an individual frame as input thus acquiring information about scenes/objects depicted in the video, the other is the temporal flow with the motion across the frames as input therefore conveying the movement of the objects; and then, a late fusion of both flows gives the final recognition result. This architecture is successful because the two streams are complementary since both flows can do action recognition on their own. But this is not consistent with our situation, as the temporal stream which usually uses optical flow fields as inputs cannot get any information about clothes and naturally excludes the ability to classify them. Therefore, the two-stream structure does not fit this joint clothes and action classification task.

Moreover, in an indoor environment, it is rare that the occupants frequently (in seconds) change their clothes or actions, indicating that a dense frame-by-frame prediction is unnecessary. While this does not contradict the fact that a fusion of predicted results on several sparsely sampled frames along time to grasp long-term information can decrease the error rate effectively. That's why we predict the category based on K (in the experiments 6 is proved to be the best value) frames around every 3.5 seconds.

Based on the above analyses and inspired by the TSNs, we implemented a single-stream network shown in Fig. A.2, where a thermal video is first divided into K segments of the same duration and then a single frame is randomly sampled from each segment as the "input" of the backbone network Inception v2, namely the GoogLeNet with Inception module (to decrease the number of parameters) and batch normalization (to accelerate convergence) [18, 19]. The output of Inception v2 is a list of prediction scores of all classes for the current frame; takes K segments into consideration, there are K lists of scores describing the original video; these lists are then calculated by an evenly average to get a sole score list which is finally fed into a 15-category Softmax classifier layer giving the predicted class label.

4. Experiments

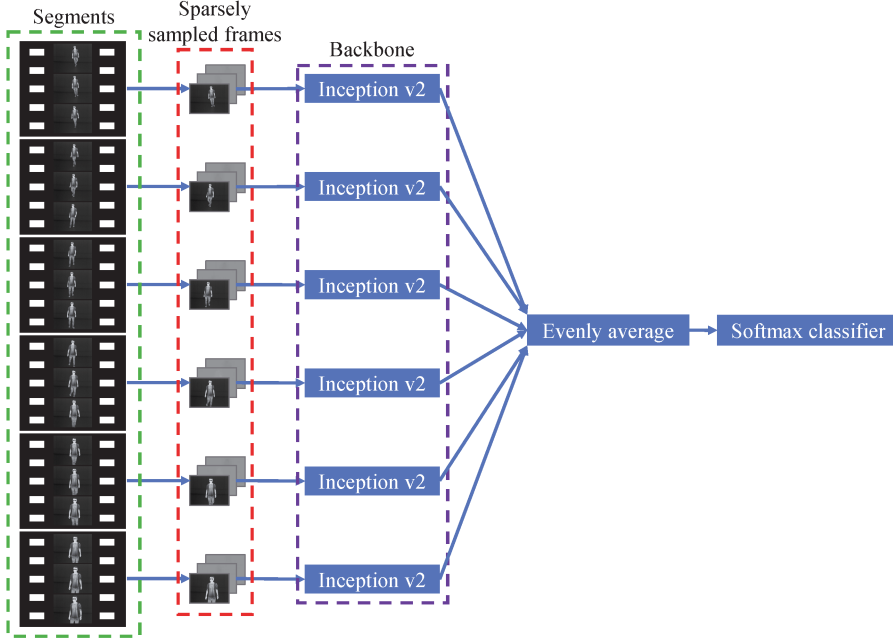


Fig. A.2: Network for clothes and action classification.

It is to be noted that the word “input” in the previous paragraph is with a quotation mark since it has several modalities: (i) an original 24-bit thermal image; (ii) a concatenated image of an 8-bit thermal image (the first channel of the original 24-bit one) and its two optical flow fields (in horizontal and vertical directions) extracted by the TVL1 algorithm [20].

4 Experiments

In this section, we first introduce the collected thermal dataset; then we test the proposed method.

4.1 Dataset Information

As there is no public thermal dataset for clothes classification, we collected a dataset in September 2019 when the average indoor temperature is around 24°C (measured by Rosenberg thermometer 66762), which gave a wide range of garment choices for occupants to select. 20 subjects (4 females, 16 males) were asked to stand, sit and walk in front of the thermal camera with at most 5 clothes types (t-shirts, long sleeves, rolled-up sleeves, long sleeves and unzipped zipper, rolled-up sleeves and unzipped zipper). We regulate the 3

actions and 5 garment categories as they are the most usual cases in an office/classroom environment, besides, a person with his/her sleeves rolled up or zipper unzipped is an immediate signal of hot feelings in most situations. When recording videos, we encouraged each subject to behave as naturally as usual, therefore, folded arms, cross legs, akimbo pose and other spontaneous postures existed in the dataset, and we kept each original video as long as 30 seconds or above allowing us to divide it into short videos with durations about 3.5 seconds thus increasing the amount of videos greatly. In this way, we obtained 291 long videos which were then trimmed into 2422 short videos. For a clear description, Fig. A.3(a), (b) and (c) illustrate the sampled frames of SitLongUnzip (sit with long sleeves and unzipped zipper), StandRollUnzip (stand with rolled-up sleeves and unzipped zipper) and WalkTshirts (walk with t-shirts), respectively. Table A.3 gives detailed number of short videos in the 15 categories. There is indeed a data unbalance problem due to two reasons: (i) not every subject wore all the 5 clothes types; (ii) only in an approximately front-view can a short video of walking be used for clothes classification.

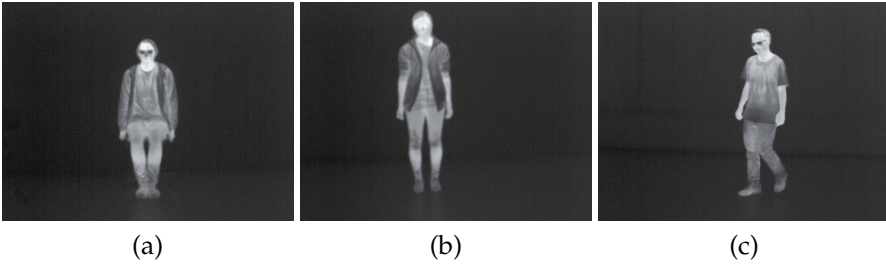


Fig. A.3: Sampled frames of SitLongUnzip (a), StandRollUnzip (b) and WalkTshirts (c).

Table A.3: Number of short videos in 15 categories.

-	Stand	Sit	Walk
T-shirts	114	125	47
Long Sleeves	284	308	109
Rolled-up Sleeves	279	309	115
Long Sleeves and Unzipped Zipper	157	155	59
Rolled-up Sleeves and Unzipped Zipper	144	161	56

When separating the dataset into the training set and testing set, we follow the principle of impartiality strictly, which means the subjects of the test set and the subjects of the training set should be totally different thus giving no biased clue for the testing phase to refer. This separation results in 4

4. Experiments

partitions (4-fold cross validation with 5 subjects corresponding to 1 fold) of training/testing sets in Table A.4.

Table A.4: Training/testing sets (S.V.N. means short video number).

-	Training Set		Testing Set	
-	Subjects ID	S.V.N	Subjects ID	S.V.N
Partition 1	1-15	1683	16-20	739
Partition 2	1-10, 16-20	1810	11-15	612
Partition 3	1-5, 11-20	1938	6-10	484
Partition 4	6-20	1835	1-5	587

4.2 Evaluation of the Proposed Method

Implementation Details

With Ubuntu 16.04 LTS, Python 3.5.2, PyTorch 1.2.0, CUDA 9.2, one NVIDIA GeForce RTX 2080 Ti as the software and hardware platform, the whole network is finetuned with the mini-batch stochastic gradient descent to learn the weights specific for the collected dataset under the basis of the backbone Inception v2 initialized with a pretrained model from ImageNet [21]. The learning rate is initialized as 0.001 and then decreases to its 1/10 after every 2,000 iterations (35 epochs), and the whole training ceases after 120 epochs.

To avoid overfitting, data augmentation is employed. Multi-scale cropping which makes the height and width of the cropped region randomly selected from $\{100\%, 90\%, 80\%, 75\%\}$ of the frame shorter side and horizontal flipping enlarge the amount and diversity of the training set effectively. This crop strategy guarantees that all the cropped regions can always have the subject located in them. Finally, all the original training set as well as the augmented ones will be resized to 224×224 as the network requires. Besides, a dropout ratio of 0.8 and a weight-decay parameter of 0.1 are also set as regularization terms to improve the network's generalization ability.

Results of Different Input Modalities

Here we evaluate the influences of different input modalities (see the last paragraph in Section 3) on the classification accuracy.

An experiment is first done with partition 1 and the results on the testing set are illustrated in Table A.5. From it, the 24-bit thermal image modality corresponds to low performance as this kind of input conveys no information of human motions along time. However, the 8-bit thermal image and optical flows -v1 modality unexpectedly works worse. A reasonable explanation is

that even though this modality combines both the spatial and temporal information, the too high frame rate at 90 fps determines the fact that almost no changes exist within two adjacent frames, therefore, the extracted temporal fields are empty. This explanation is well supported by the improved performances of v2 and v3 where the frame rates are 45 fps and 30 fps, respectively.

Table A.5: Classification accuracy with different input modalities (CAoTS means classification accuracy on testing set).

Modality	Partition	CAoTS
24-bit thermal image	1	83.09%
8-bit thermal image and optical flows -v1	1	82.14%
8-bit thermal image and optical flows -v2	1	84.57%
8-bit thermal image and optical flows -v3	1	94.32%
8-bit thermal image and optical flows -v3	2	87.91%
8-bit thermal image and optical flows -v3	3	96.28%
8-bit thermal image and optical flows -v3	4	94.21%

Based on the above comparison, we fix the modality as 8-bit thermal image and optical flows -v3, and use the remaining partitions to further assess the method. We achieve performances of 87.91%, 96.28%, 94.21% on partition 2, partition 3, partition 4, respectively, also listed in Table A.5.

Result of an Input Preprocessing

The drop in performance on partition 2 stands out. By observing its confusion matrix (see Fig. A.4) and comparing the predicted labels with ground truth, we find that the incorrectly classified cases only relate to the clothes type and mainly belong to two subjects (ID 12 and ID 15). A common problem of these wrongly classified videos is that the contrast in the region containing the person is low due to the overall high brightness, thus raising the difficulty of distinguishing clothing from skin, not to mention the clothes type classification. To solve this issue, a linear grayscale transformation for the foreground region is done by stretching its original grayscale distribution to an 8-bit range $[0, 255]$ thus increasing the subtle gray value difference between human skin and clothing to some extent. This input preprocessing for test videos from ID 12 and ID 15 boosts the classification accuracy on partition 2 to 95.75% with nothing else changed, and the new confusion matrix is shown in Fig. A.4(b). From Fig. A.4(b), the inability to differentiate long sleeves/rolled-up sleeves from t-shirts occurred in Fig. A.4(a) is successfully solved, but a new problem of SitRoll-StandRoll confusion caused by the disappearance of the chair after processing stresses the importance of selecting a proper image preprocessing. The original frame (Fig. A.5(a)), its

4. Experiments

StandTshirts	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
StandLong	10	54	0	0	0	0	0	0	0	0	0	0	0	0	0
StandLongUnzip	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0
StandRoll	16	0	0	47	0	0	0	0	0	0	0	0	0	0	0
StandRollUnzip	0	0	1	0	34	0	0	0	0	0	0	0	0	0	0
SitTshirts	0	0	0	0	0	35	0	0	0	0	0	0	0	0	0
SitLong	0	0	0	0	0	14	64	0	0	0	0	0	0	0	0
SitLongUnzip	0	0	0	0	0	1	0	41	0	0	0	0	0	0	0
SitRoll	0	0	0	0	0	18	0	0	53	0	0	0	0	0	0
SitRollUnzip	0	0	0	0	0	0	0	0	0	49	0	0	0	0	0
WalkTshirts	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0
WalkLong	1	0	0	0	0	0	0	0	0	0	3	27	0	0	0
WalkLongUnzip	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0
WalkRoll	0	0	0	1	0	0	0	0	0	0	9	0	0	22	0
WalkRollUnzip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18
	StandTshirts	StandLong	StandLongUnzip	StandRoll	StandRollUnzip	SitTshirts	SitLong	SitLongUnzip	SitRoll	SitRollUnzip	WalkTshirts	WalkLong	WalkLongUnzip	WalkRoll	WalkRollUnzip

(a)

StandTshirts	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
StandLong	1	65	0	0	0	0	0	0	0	0	0	0	0	0	0
StandLongUnzip	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0
StandRoll	3	1	0	59	0	0	0	0	0	0	0	0	0	0	0
StandRollUnzip	0	0	1	0	34	0	0	0	0	0	0	0	0	0	0
SitTshirts	0	0	0	0	0	35	0	0	0	0	0	0	0	0	0
SitLong	0	0	0	0	0	0	78	0	0	0	0	0	0	0	0
SitLongUnzip	0	0	0	0	0	1	0	41	0	0	0	0	0	0	0
SitRoll	0	0	0	8	0	0	0	0	63	0	0	0	0	0	0
SitRollUnzip	0	0	0	0	0	0	0	0	0	49	0	0	0	0	0
WalkTshirts	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0
WalkLong	1	0	0	0	0	0	0	0	0	0	3	27	0	0	0
WalkLongUnzip	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0
WalkRoll	0	0	0	5	0	0	0	0	0	0	2	0	0	25	0
WalkRollUnzip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18
	StandTshirts	StandLong	StandLongUnzip	StandRoll	StandRollUnzip	SitTshirts	SitLong	SitLongUnzip	SitRoll	SitRollUnzip	WalkTshirts	WalkLong	WalkLongUnzip	WalkRoll	WalkRollUnzip

(b)

Fig. A.4: Confusion matrices before and after a preprocessing on partition 2. (a) Before the preprocessing. (b) After the preprocessing.

foreground mask (Fig. A.5(b)) and the grayscale-stretched frame (Fig. A.5(c)) of an incorrectly classified case from subject ID 12 are displayed below. From Fig. A.5(c), it is easy to see that the foreground region has more details indicating the enhanced distinction between the human skin and her clothing compared with that in Fig. A.5(a), which explains the big improvement of the classification rate.

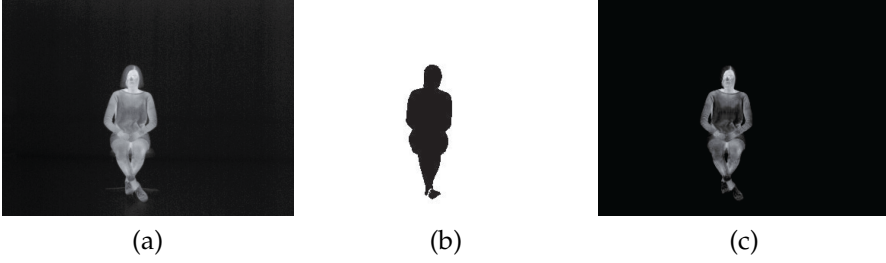


Fig. A.5: The original frame (a), foreground mask (b) and preprocessed frame (c) of an incorrectly classified case from Subject ID 12.

Good Practices

From the above evaluation of the proposed method, we summarize some good practices for this sort of task: (i) the modality fusing 8-bit spatial image and two temporal flows is an effective way to tackle video-based classification problems; (ii) temporal information extracted between two adjacent frames is closely related to the video's frame rate; (iii) an appropriate frame preprocessing scheme to enhance the image details is quite useful for the following predictions. With all these good practices, the proposed method finally achieves an average classification rate of 95.14% on 4 partitions (separate rates of 94.32%, 95.75%, 96.28% and 94.21% on partition 1, partition 2, partition 3 and partition 4, respectively).

5 Conclusions

In this paper, we proposed a vision-based individual factor acquisition method relying on a CNN classifier to acquire an occupant's clothes and action type in a built environment, which can be used to estimate CLO and MET thus assessing his/her thermal comfort levels. A new thermal dataset with 20 subjects verifies the feasibility and effectiveness of the proposed method with an average 15-category classification accuracy at 95.14% on 4 dataset partitions. These results prove the value of this computer-vision research in personalized thermal comfort assessment and the further potential

in urban design and energy saving. Future research will take more clothes and action types into consideration for a multi-person scenario.

References

- [1] P. O. Fanger, "Thermal comfort. analysis and applications in environmental engineering," *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.
- [2] Chapter, Thermal Comfort, "Fundamentals Volume of the ASHRAE Handbook. ASHRAE." Inc., Atlanta, GA., 2005.
- [3] Q. Zhao, Y. Zhao, F. Wang, J. Wang, Y. Jiang, and F. Zhang, "A data-driven method to describe the personalized dynamic thermal comfort in ordinary office environment: From model to application," *Building and Environment*, vol. 72, pp. 309–318, 2014.
- [4] A. Ghahramani, C. Tang, and B. Becerik-Gerber, "An online learning approach for quantifying personalized thermal comfort via adaptive stochastic modeling," *Building and Environment*, vol. 92, pp. 86–96, 2015.
- [5] F. Auffenberg, S. Stein, and A. Rogers, "A personalised thermal comfort model using a bayesian network," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [6] S. Sim, M. Koh, K. Joo, S. Noh, S. Park, Y. Kim, and K. Park, "Estimation of thermal sensation based on wrist skin temperatures," *Sensors*, vol. 16, no. 4, p. 420, 2016.
- [7] Z. Wu, N. Li, H. Cui, J. Peng, H. Chen, and P. Liu, "Using upper extremity skin temperatures to assess thermal comfort in office buildings in changsha, china," *International journal of environmental research and public health*, vol. 14, no. 10, p. 1092, 2017.
- [8] T. Chaudhuri, D. Zhai, Y. C. Soh, H. Li, and L. Xie, "Thermal comfort prediction using normalized skin temperature in a uniform built environment," *Energy and Buildings*, vol. 159, pp. 426–440, 2018.
- [9] A. Ghahramani, G. Castro, B. Becerik-Gerber, and X. Yu, "Infrared thermography of human face for monitoring thermoregulation performance and estimating personal thermal comfort," *Building and Environment*, vol. 109, pp. 1–11, 2016.
- [10] A. Ghahramani, G. Castro, S. A. Karvigh, and B. Becerik-Gerber, "Towards unsupervised learning of thermal comfort using infrared thermography," *Applied Energy*, vol. 211, pp. 41–49, 2018.

- [11] H. Na, J.-H. Choi, H. Kim, and T. Kim, "Development of a human metabolic rate prediction model based on the use of kinect-camera generated visual data-driven approaches," *Building and Environment*, p. 106216, 2019.
- [12] X. Cheng, B. Yang, T. Olofsson, G. Liu, and H. Li, "A pilot study of online non-invasive measuring technology based on video magnification to determine skin temperature," *Building and Environment*, vol. 121, pp. 1–10, 2017.
- [13] D. Li, C. C. Menassa, and V. R. Kamat, "Robust non-intrusive interpretation of occupant thermal comfort in built environments with low-cost networked thermal cameras," *Applied Energy*, vol. 251, p. 113336, 2019.
- [14] S. Lu, C. Hameen, and A. Aziz, "Dynamic hvac operations with real-time vision-based occupant recognition system," in *2018 ASHRAE Winter Conference, Chicago*, 2018.
- [15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.
- [16] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 363–378.
- [17] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.
- [18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [20] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, pp. 214–223.
- [21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

Paper B

Automatic Estimation of Clothing Insulation Rate and Metabolic Rate for Dynamic Thermal Comfort Assessment

Jinsong Liu, Isak Worre Foged, and Thomas B. Moeslund

The paper has been published in the journal of
Pattern Analysis and Applications

© 2021 The Authors, under exclusive licence to Springer-Verlag London Ltd.
part of Springer Nature 2021

First published in Jinsong Liu, Isak Worre Foged, and Thomas B. Moeslund.
“Automatic estimation of clothing insulation rate and metabolic rate for
dynamic thermal comfort assessment.” *Pattern Analysis and Applications*
(2021). <https://doi.org/10.1007/s10044-021-00961-5>

Reproduced with permission from Springer Nature.

The layout has been revised.

Abstract

Existing heating, ventilation, and air-conditioning (HVAC) systems have difficulties in considering occupants' dynamic thermal needs, thus resulting in overheating or overcooling with huge energy waste. This situation emphasizes the importance of occupant-oriented microclimate control where dynamic individual thermal comfort assessment is the key. Therefore, in this paper, a vision-based approach to estimate individual clothing insulation rate (I_{cl}) and metabolic rate (M), the two critical factors to assess personal thermal comfort level, is proposed. Specifically, with a thermal camera as the input source, a convolutional neural network (CNN) is implemented to recognize an occupant's clothes type and activity type simultaneously. The clothes type then helps to differentiate the skin region from the clothing-covered region, allowing to calculate the skin temperature and the clothes temperature. With the two recognized types and the two computed temperatures, I_{cl} and M can be estimated effectively. In the experimental phase, a novel thermal dataset is introduced, which allows evaluations of the CNN-based recognizer module, the skin and clothes temperatures acquisition module, as well as the I_{cl} and M estimation module, proving the effectiveness and automation of the proposed approach.

keywords: thermal camera; computer vision; clothing insulation rate; metabolic rate; thermal comfort

1 Introduction

In modern life, people spend more time indoors than ever before, especially for office workers. To maintain a comfortable indoor environment thus improving their working efficiency, heating, ventilation, and air-conditioning (HVAC) facilities operate continuously, consuming nearly half of the indoor used energy according to the residential energy consumption survey in 2005 [1].

Almost all HVAC systems are temperature-aimed so that occupants can set empirical temperatures like 22°C in winter and 26°C in summer. However, this ignores personal dynamic thermal sensations and needs, which results in overheating or overcooling with huge energy waste.

To address this problem, dynamic microclimate control by classical HVAC systems or emerging deformable envelopes to change solar radiation is important, in which human thermal comfort assessment is critical. To facilitate applications by HVAC industry, Fanger [2] first presented a predicted mean vote (PMV) model with 7-scaled thermal sensations (see Table B.1) to assess occupants' average thermal comfort level in a laboratory environment. Accordingly, this model recommends a very narrow range ($-0.5 \leq \text{PMV} \leq +0.5$) representing that at least 90% of the occupants are satisfied with a certain

environment. Later for its validity and applicability in the industry, the PMV model was recognized as a standard in ISO 7730 [3] and widely applied to more controlled indoor microclimates [4–9].

Table B.1: Seven-scale list of thermal sensations [2].

Sensation	Scale
Hot	3
Warm	2
Slightly warm	1
Neutral	0
Slightly cool	-1
Cool	-2
Cold	-3

In detail, the PMV model is determined by six factors, that is, four environmental values (air temperature (t_a), mean radiation temperature (\bar{t}_r), relative humidity (RH), air velocity (V_a)) and two personal values (clothing insulation rate (I_{cl}), metabolic rate (M)). I_{cl} means the thermal insulation ability provided by clothing to protect the skin from a cold or hot environment outside the clothes, indicating a person's heat conduction and transfer. M means the amount of energy used by a person per unit of time, indicating his/her self heat generation. In practical applications, t_a , \bar{t}_r , RH , and V_a can be measured by the thermometer, hygrometer, and anemograph. The acquisition of I_{cl} and M requires either large manual work or expensive precise instruments. Therefore, realizing accurate and dynamic thermal comfort assessment is still difficult. To improve this situation and further help to build a thermal adaptive architecture, we propose an automatic estimation scheme of I_{cl} and M from extracted key factors (the clothes type, the activity type, the skin temperature, and the clothes temperature), thus facilitating Fanger's model to be applied to modern smart buildings. The concrete contributions are:

1. A vision-based contactless method to automatically estimate both I_{cl} and M for a single-person scenario is proposed, which is also privacy-preserving as a low-resolution (384×288) thermal camera is the only input source.
2. A convolutional neural network (CNN) considering both spatial and temporal information to recognize an occupant's clothes type and activity type simultaneously is designed, achieving an average classification rate of 95.17% on 6 test partitions for a 15-category scenario.
3. The feasibility of applying OpenPose [10] to the thermal mode is experimentally and quantitatively verified.

2. Related Work

4. A comprehensive I_{cl} estimation model considering multiple factors (the clothes type, the activity type, the skin temperature, and the clothes temperature) is suggested.
5. The M of an occupant is conveniently estimated from his/her activity type.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the proposed method. Section 4 gives the experiments. Section 5 concludes this paper and discusses future work.

2 Related Work

For I_{cl} estimation, the published works can be classified into clothes-unrelated methods and clothes-related methods according to whether the method correlates to the specific clothes the occupant wears or not.

Among clothes-unrelated methods, the simplest way is to set a fixed I_{cl} (1.0 *clo* in winter, 0.5 *clo* in summer, 1 *clo* = 0.155 m²K/W) on accordance with the ISO 7730 [3] or ASHRAE Standard 55 [11]. Other works found the correlation between clothing insulation levels and the season [12], outdoor temperature [12–17], indoor temperature [13, 14], gender [12, 14], latitude [14], mode of transportation [12], and so on, and then used the relationship models (linear or nonlinear regression) to estimate the I_{cl} . Such methods ignore the intrinsic properties of clothes and usually need time-consuming questionnaires, which is why some researches move towards the clothes-related methods.

The typical tool in clothes-related methods is the thermal manikin dressed in the garments to be estimated and placed in an environmental chamber [18]. But one research [19] pointed out that the I_{cl} values measured on a thermal manikin and volunteers are different, and ISO 9920 [18] mentioned that the cost for the specialized equipment based on manikins is beyond the reach of most people. Therefore more researches did I_{cl} measurements directly with humans. Researches [19, 20] used ordinary infrared sensors attached to the subject to calculate I_{cl} from the temperature difference between the skin and the clothes. Works [21–23] calculated I_{cl} simply from measuring the weight of the clothes. These two types of approaches are low-cost but they are inconvenient due to their interference with subjects. Hence contactless methods started to appear. Lee in 2016 [24], Miura in 2019 [25], and Lee in 2020 [26] all used an infrared camera to measure the skin and clothes temperatures to estimate I_{cl} . However, [24] and [26] failed to give information about how to acquire temperatures of clothes and skin automatically, making them difficult to be applied without manual work. [25] and [26] did not consider the influence of activities (sitting, standing, or walking) on I_{cl} . And all these

three researches ignored the I_{cl} changes because of rolled-up sleeves and unzipped zippers that occur frequently in daily life. As a whole, automatic and comprehensive estimation of I_{cl} is still underexploited.

For M estimation, most works rely on specific equipment, for the reason that M could be calculated from a person's oxygen (O_2) consumption and carbon dioxide (CO_2) generation [27–29], heart rate (HR) [30–33], breathing frequency (BF) [30, 31], or blood pressure (BP) [34]. The Vmax Encore Metabolic Cart used by Luo [27] and the COSMED K5 used by Zhai [28] can measure O_2 and CO_2 values simultaneously, and both the equipment need masks that cover the subject's nose and mouth. The Telariire-7001 and the Philips D12 used by Ji [29] are handheld devices measuring the amounts of inhaled O_2 and exhaled CO_2 separately. Calvaresi [30] and Casaccia [31] utilized a BioHarness 3.0 attached on the chest to get a person's HR and BF. Na [32] and Hasan [33] used wearable Fitbit smart bracelets to measure the HR. Gilani [34] employed an arm blood pressure monitor to obtain BP. Even though equipment-based M estimation methods are relatively accurate, a common problem is that the devices are skin-attached, which not only causes inconvenience but also remains impossible to use outside a laboratory. On the other hand, ISO 8996 [35], ISO 7933 [36], ISO 7730, and ASHRAE 55 all give standard M values for specific activities (see Table B.2), indicating an efficient estimation approach, and thus work [32] used this as M reference values with the person's activity known as a manual recording knowledge.

Table B.2: Metabolic rate for specific activities [35].

Activity	M (W/m ²)
Sleeping	40
At rest, sitting	55
At rest, standing	70
Walking on the level, even path, solid	
1. Without load	
At 2 km/h	110
At 3 km/h	140
At 4 km/h	165
At 5 km/h	200
2. With load	
10 kg, 4 km/h	185
30 kg, 4 km/h	250

Table B.3 summarizes the main information of the studied literature. It is obvious to see that current I_{cl} and M estimation methods rely on either inconvenient and expensive devices or time-consuming manual work.

As the specific work about vision-based I_{cl} or M estimation is very lim-

2. Related Work

Table B.3: Main information of the studied literature about I_{cl} and M estimation.

Type	Source	Approach	Drawback
I_{cl}	[3, 11]	Fixed values: 1.0 <i>clo</i> in winter, 0.5 <i>clo</i> in summer.	Not dynamic.
I_{cl}	[12–17]	Estimation from seasons, outdoor and indoor temperatures, gender, etc.	Ignore clothes properties, require questionnaires.
I_{cl}	[19, 20]	Estimation from skin and clothes temperatures obtained by attached sensors.	Interference with subjects.
I_{cl}	[21–23]	Estimation from clothes weight.	Interference with subjects.
I_{cl}	[24–26]	Estimation from skin and clothes temperatures obtained by a thermal camera.	Require manual work, ignore some factors like activities and clothes types.
M	[27–29]	Estimation from O_2 consumption and CO_2 generation.	Require expensive equipment, inconvenient masks.
M	[30–34]	Estimation from heart rate, breathing frequency, or blood pressure.	Require specific equipment, inconveniently wearable.
M	[3, 11, 32, 35, 36]	Estimation from activities.	Require manual work.

ited, to broadly investigate the developments of the thermal comfort field relying on computer vision, more surveys are done. Cheng [37, 38] used a normal RGB camera to take pictures of the hand back and then investigated the relationship between the hand back skin color saturation and the skin temperature. Jazizadeh [39] also used an RGB camera to take pictures of facial areas and then assessed human thermal comfort levels by analysing face illumination differences caused by blood flow variations. Li [40, 41] used a thermal camera to detect facial skin temperatures and then mapped the temperatures into three personal thermal feelings of hot, neutral, and cold. Lu [42] sought help from machine learning (ML) algorithms of random forest (RF) and support vector machine (SVM) to predict thermal sensations from a feature set consisting of indoor air temperature, relative humidity, skin temperature, and clothing surface temperature, where the temperatures were acquired from a thermal camera. Qian [43] studied the direct correlation between a person's pose (shivering, hand rubbing, etc.) with his/her thermal

feelings (hot or cold) by using an RGB camera as the data source.

To sum up, no matter contacted or contactless methods for estimating I_{cl} , most of them require labor-some manual work; methods for M estimation have to use expensive devices; vision-based researches on thermal comfort assessment usually take advantage of new ML models instead of Fanger’s model to directly predict thermal sensations from skin temperatures, which ignores many useful environmental/personal factors. Different from the related works, the proposed method has such characteristics: (1) it can estimate both I_{cl} and M ; (2) the I_{cl} estimation considers multiple factors including skin temperature, clothes temperature, clothes type, and activity type, which is more comprehensive; (3) the estimations of both I_{cl} and M rely on computer vision algorithms without requiring extra manual work, which is automatic; (4) the estimated I_{cl} and M are used as personal factors in Fanger’s model (well proved across years) for dynamic thermal comfort assessment, which is solid and reliable.

3 Proposed Method

In this section, we describe our approach. The idea which is illustrated in Fig. B.1 consists of three components:

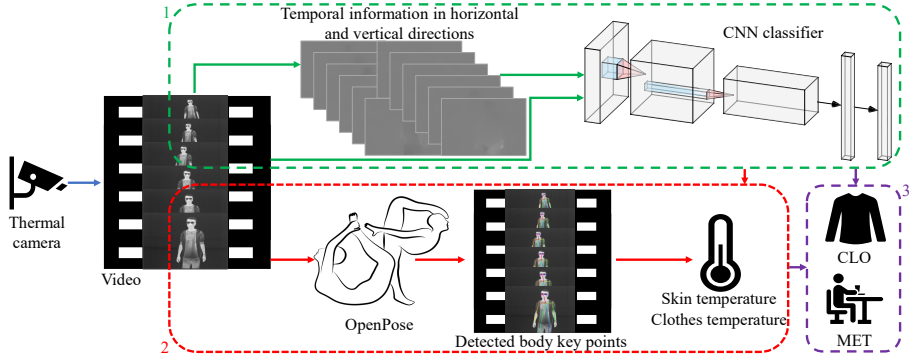


Fig. B.1: Overview of the proposed method.

1. A CNN classifier is implemented to recognize an occupant’s clothes type and activity type at the same time. See the green dashed box in Fig. B.1.
2. The occupant’s body key points (like the nose, neck, shoulders, elbows, and so on) are detected by OpenPose [10], which act as body part location references. These body part landmarks together with the recognized clothes type help to differentiate the skin-bare region from the

clothing-covered region, and thus the skin temperature and the clothes temperature are calculated. See the red dashed box in Fig. B.1.

3. A basic I_{cl} is calculated from the skin temperature and the clothes temperature, and then the final I_{cl} is calculated by additional corrections involved with the clothes type and activity type. M is estimated from the recognized activity type directly. See the purple dashed box in Fig. B.1.

In the following contents, details of each component are provided.

3.1 Clothes and Activity Recognition

Clothes type and activity type can be recognized via two phases, but we prefer to predict both tasks at the same time, which is more efficient and practically useful. However, two-stream CNNs [44–46] cannot realize this expectation as the temporal information is usually represented by optical flows that have no information about clothes. Hence, we simplify the two-stream network into a single stream network with its three-channel input containing both spatial information (represented by one 8-bit thermal frame) and temporal information (represented by two 8-bit optical flows in horizontal and vertical directions calculated by the TVL1 algorithm [47]). Moreover, in an indoor environment, an occupant seldom changes his/her clothes and activity very frequently as the frame rate, indicating that a frame-by-frame recognition is unnecessary. But a fusion of predictions on several frames across time like [46] did is still very important, as this can capture essential long-term information to reduce the classification error rate. [48]

Based on the above analysis we extend our previous work [48]. Concretely, we implement a CNN with its architecture shown in Fig. B.2 [48]. In the CNN, a thermal video is first divided into K segments of the same length, and then a frame is randomly sampled from each segment as the “input” of the backbone network Inception v2 [49, 50]; the output of the backbone is a list of prediction scores of all classes for the current “input”; because of K segments, there are K lists of scores describing the original thermal video; these lists are then evenly averaged as a single score list that is fed into a softmax classifier layer to give the predicted class label [48].

It is to be noted that the word “input” in the last paragraph is with a quotation mark as it means a concatenated image composed of one 8-bit thermal frame and its corresponding two 8-bit optical flows [48]. And the implemented CNN is a 15-category classifier indicating five clothes types and three activity types that are the primary types in a normal office environment (see Table B.4 [48]).

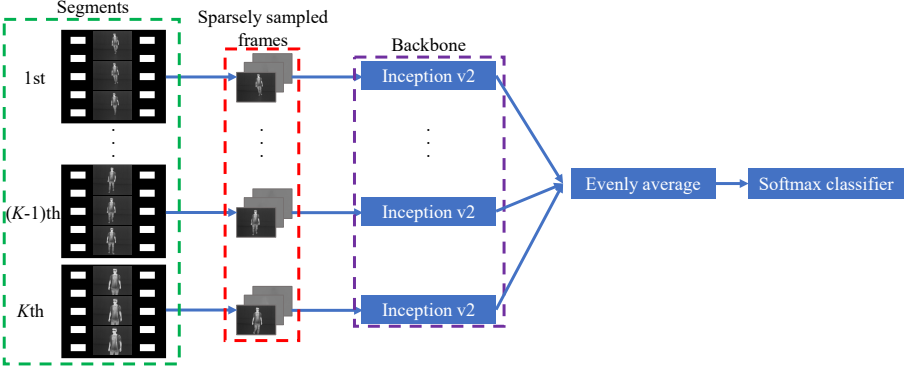


Fig. B.2: Network for clothes and action recognition.

Table B.4: Details of the 15 categories.

Category	Meaning
SitLong	Long-sleeved clothes, sitting
SitLongUnzip	Long-sleeved and zipper unzipped clothes, sitting
SitRoll	Sleeves rolled-up clothes, sitting
SitRollUnzip	Sleeves rolled-up and zipper unzipped clothes, sitting
SitTshirts	Short-sleeved t-shirt, sitting
StandLong	Long-sleeved clothes, standing
StandLongUnzip	Long-sleeved and zipper unzipped clothes, standing
StandRoll	Sleeves rolled-up clothes, standing
StandRollUnzip	Sleeves rolled-up and zipper unzipped clothes, standing
StandTshirts	Short-sleeved t-shirt, standing
WalkLong	Long-sleeved clothes, walking
WalkLongUnzip	Long-sleeved and zipper unzipped clothes, walking
WalkRoll	Sleeves rolled-up clothes, walking
WalkRollUnzip	Sleeves rolled-up and zipper unzipped clothes, walking
WalkTshirts	Short-sleeved t-shirt, walking

3.2 Skin and Clothes Temperatures Acquisition

A person's temperature distribution is the most direct way to indicate his/her thermal sensation, what's more, I_{cl} can be effectively estimated from the skin temperature and the clothes temperature.

To calculate both temperatures, we first utilize the key points detected by OpenPose as location references of body parts. We focus on the I_{cl} of the upper body, and hence detect the right eye, left eye, nose, neck, right shoulder, left shoulder, right elbow, left elbow, right wrist, and left wrist. Besides, for more precise body parts localization, we also calculate the coordinates of the middle points between the nose and the neck, the neck and the shoulders, the

3. Proposed Method

shoulders and the elbows, respectively, with the help of the detected points.

Then, the clothes type predicted by the above CNN helps to distinguish the skin region R_s from the clothes region R_c . That is, for clothes with sleeves rolled-up or t-shirt with short sleeves, the lower arm region is treated as the skin region R_s , while for clothes with long sleeves, the lower arm region belongs to the clothes region R_c . Besides, as the face area and the chest area are the critical segments representing a person's temperature distribution [51, 52], they are classified as the skin region R_s and the clothes region R_c , respectively.

For a better explanation, Fig. B.3 shows the detailed skin region R_s and clothes region R_c of a standing person wearing long-sleeved clothes and sleeves rolled-up clothes, in which the red hollow circles represent the key points detected by OpenPose; the blue crosses are the additional middle points of two nearby key points; the green box and green lines mean the skin region R_s ; the yellow box and yellow lines mean the clothes region R_c . Specifically, the green region is the segment below the nose while above the jaw with the vertical coordinates of two eyes as the right and left boundaries, thus avoiding the influences of glasses and the neck covered by a collar. The yellow region is the segment below the neck while above the elbow with proper vertical boundaries to exactly cover the chest.

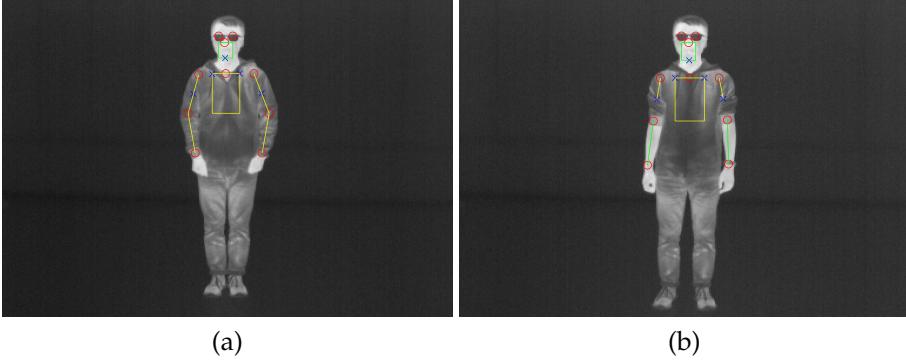


Fig. B.3: Skin region and clothes region of a standing person in two different garments. (a) Long-sleeved clothes. (b) Sleeves rolled-up clothes.

As a whole, the skin region R_s and the clothes region R_c are summarized in Table B.5, where the specific coordinates of the regions are obtained with the help of key points from OpenPose mentioned above.

At last, the skin temperature T_s and the clothes temperature T_c can be calculated as the average value from the skin region R_s and the clothes region R_c , respectively.

$$T_s = \frac{\sum_{(x,y) \in R_s} T_{x,y}}{\sum_{(x,y) \in R_s} 1} \quad (\text{B.1})$$

Table B.5: Skin region R_s and clothes region R_c .

Clothes Category	Skin Region R_s	Clothes Region R_c
Long-sleeved clothes	Face	Chest & arms
Sleeves rolled-up clothes & t-shirt	Face & lower arms	Chest & upper arms

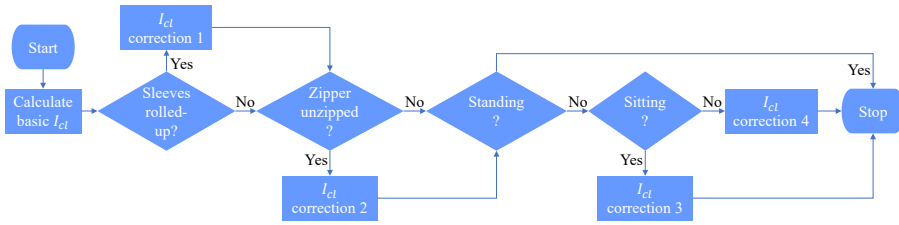
$$T_c = \frac{\sum_{(x,y) \in R_c} T_{x,y}}{\sum_{(x,y) \in R_c} 1} \quad (\text{B.2})$$

here $T_{x,y}$ is the temperature at location (x, y) which can be acquired directly from the thermal camera. In our case, the utilized thermal camera (Xenics-Gobi-384) is a thermography camera that can give the temperature of every location in the captured scene. In addition, it has a NETD (noise-equivalent temperature difference) of 80mK in the temperature range from -20°C to 120°C , which means a tiny temperature change of 0.08°C can be captured by the camera. Therefore, this camera is a satisfactory temperature acquisition device for skin and clothes.

3.3 I_{cl} and M Estimation

This part utilizes the clothes type, the activity type, the skin temperature T_s , and the clothes temperature T_c , extracted from Section 3.1 and Section 3.2, to estimate I_{cl} (in unit *clo*) and M (in unit W/m^2).

The steps of I_{cl} estimation are illustrated in Fig. B.4, where the corrections 1 to 4 are based on ISO 9920 [18], and the detailed derivations of corresponding Equation B.4 to Equation B.8 are in the appendix.

**Fig. B.4:** Diagram of estimating I_{cl} .

Firstly, a basic clothing insulation rate $B_{I_{cl}}$ is calculated from T_s and T_c by Equation B.3 [25]:

$$B_{I_{cl}} = \frac{1}{0.155 \cdot h} \cdot \frac{T_s - T_c}{T_c - T_o} \quad (\text{B.3})$$

where 0.155 is the scale of the unit-conversion from $\text{m}^2\text{K}/\text{W}$ to *clo*; h is the heat transfer coefficient of a human which is set to 8.6 [25]; T_o is the operative

3. Proposed Method

temperature which takes into account not only the indoor air temperature but also the radiant temperatures of surfaces (floor, walls, and ceiling); therefore, we use the average of the temperatures of the background scene which excludes the human foreground to represent T_o . It is to be noted that with T_s and T_c extracted from our method, any calculation model for estimating the basic clothing insulation rate $B_{I_{cl}}$ can be used.

Secondly, according to the person's clothes type, some corrections may happen, that is, I_{cl} correction 1 and I_{cl} correction 2, see Fig. B.4.

I_{cl} correction 1 If the sleeves of the worn clothes are rolled up, the person's lower arms are not covered by clothing, resulting in a decrease in clothing insulation; hence, the basic I_{cl} value $B_{I_{cl}}$ is corrected to:

$$C1_{I_{cl}} = B_{I_{cl}} - 0.00874 \cdot A_{cov,0} \quad (B.4)$$

where $A_{cov,0}$ is the increased body surface area not covered by clothing. In this case, it is equal to 6.2.

I_{cl} correction 2 If the zipper of the worn clothes is unzipped, the person's chest and abdomen region will be covered by a single clothing layer under a normal indoor condition, also leading to a decrease of clothing insulation; hence, $B_{I_{cl}}$ or $C1_{I_{cl}}$ is corrected to:

$$C2_{I_{cl}} = B_{I_{cl}} - 0.00510 \cdot A_{cov,1} \quad (B.5)$$

or

$$C2_{I_{cl}} = C1_{I_{cl}} - 0.00510 \cdot A_{cov,1} \quad (B.6)$$

where $A_{cov,1}$ is the increased body surface area covered by a single clothing layer. In this case, it is equal to 16.3.

Lastly, if the person is standing, no more correction is needed. While if he/she is sitting or walking, further correction is necessary, that is, I_{cl} correction 3 or I_{cl} correction 4, see Fig. B.4.

I_{cl} correction 3 If the person is sitting, the air layer between the skin and the clothes on the back is compressed, inducing an average I_{cl} decrease by 12%, but the office chair gives extra insulation by 0.105 *clo* on average. Hence,

$$C3_{I_{cl}} = 0.88 \cdot C_{I_{cl}} + 0.105 \quad (B.7)$$

where $C_{I_{cl}}$ could be $B_{I_{cl}}$, $C1_{I_{cl}}$, or $C2_{I_{cl}}$ according to the clothes type.

I_{cl} correction 4 If the person is walking, the body motion and the wind motion increase the air exchange via clothes openings (collars, cuffs) [18]. In an office environment with air velocity smaller than 0.2 m/s and air insulation as 0.7 *clo*, when the human walking velocity is 1 m/s, the I_{cl} value can be corrected as:

$$C4_{I_{cl}} = \begin{cases} 0.9878 \cdot C_{I_{cl}}^2 + 0.091 \cdot C_{I_{cl}}, & 0 < C_{I_{cl}} \leq 0.6 \\ 0.5927 \cdot C_{I_{cl}} + 0.0546, & 0.6 < C_{I_{cl}} < 1.4 \end{cases} \quad (B.8)$$

As for M estimation, we refer to Table B.2 to get the M for sitting, standing, and walking (3.6 km/h) being 55, 70, and 150, respectively, which avoids any additional device and manual annotation.

4 Experiments

In this section, we first introduce our collected thermal dataset, and then we evaluate the proposed method.

4.1 Dataset Information

Since there is no public dataset for clothes type recognition in thermal mode, we collected a dataset of 20 subjects in an office in September 2019. The air temperature and the air humidity in the room were around 24°C and 40% (measured by a Rosenborg thermometer 66762). The 20 subjects include four females and 16 males who were standing (about 3 meters away), sitting (about 3 meters away), and walking (about 1 to 3 meters away) from a thermal camera (Xenics Gobi-384-GigE) with at most five clothes types (long sleeves, long sleeves and unzipped zipper, rolled-up sleeves, rolled-up sleeves and unzipped zipper, t-shirt). To best detect human temperatures, the emissivity of the camera is set as 0.98. The three activities and five garments were chosen as they are the most common cases in an office environment. Furthermore, a person with his/her sleeves rolled up or zipper unzipped is an immediate signal of feeling hot in most situations. We encouraged each subject to behave naturally; therefore, folded arms, akimbo pose, crossed legs, or other spontaneous postures existed in the videos. A video (of a subject with one activity in one type of garment) was recorded with a length of 30 s or more, so we could trim it into short videos with a duration of about 3.5 s. In this way, we obtained 291 long videos which were then trimmed into 2422 short videos, see Table B.6. The data unbalance problem is due to two reasons: (i) not all subjects wore the five types of clothes; (ii) only in an approximately-front view can a short video of walking be used for clothes type recognition. [48]

Table B.6: Short video number of each category [48].

-	Sit	Stand	Walk
Long sleeves	308	284	109
Long sleeves, unzipped zipper	155	157	59
Rolled-up sleeves	309	279	115
Rolled-up sleeves, unzipped zipper	161	144	56
T-shirt	125	114	47

4.2 Evaluation of the CNN for Clothes Type and Activity Recognition

Evaluation of a CNN needs a partition of training, validation, and testing set. To avoid the bias where the three sets have common subjects, we separate the collected dataset according to the subject ID. For a comprehensive evaluation, we give 6 partitions recorded in Table B.7.

Table B.7: Training/validation/testing sets information (SID means subject ID, # means the number of videos).

-	Training		Validation		Testing	
-	SID	#	SID	#	SID	#
Partition 1	1-12	1337	13-16	476	17-20	609
Partition 2	1-12	1337	17-20	609	13-16	476
Partition 3	9-20	1586	1-4	470	5-8	366
Partition 4	9-20	1586	5-8	366	1-4	470
Partition 5	1-6, 15-20	1518	7-10	443	11-14	461
Partition 6	1-6, 15-20	1518	11-14	461	7-10	443

The hardware and software platforms for implementing the CNN are one NVIDIA GeForce RTX 2080 Ti, Ubuntu 16.04 LTS, CUDA 9.2, Python 3.5.2, and PyTorch 1.2.0. The network backbone Inception v2 is pretrained with ImageNet [53] dataset, and then the whole CNN is finetuned with the collected thermal dataset. The number of segments divided from a video is 6 which is the largest value considering the used GPU capacity. The learning rate is initialized as 0.0005 and then multiply with a factor of 0.1 at the 60th epoch and the 120th epoch, respectively. The whole training ceases at the 150th epoch where the CNN has already converged.

Data augmentation strategies including a multiscale cropping and a horizontal flipping are also implemented. The height and the width of the cropped region are randomly selected from $\{288, 274, 260\}$ which refer to $\{100, 95, 90\}\%$ of the shorter edge size of the thermal frame (384×288), guaranteeing all the cropped regions always contain the subject. Finally, all the original frames as well as the augmented ones are resized to 224×224 as the network requires. Besides, to improve the network’s generalization ability, a dropout ratio of 0.75 and a weight-decay parameter of 0.005 are also used. [48]

For each partition, we use the best-trained model on the validation set to evaluate the testing set. At first, we get a poor classification result on the 6 partitions, see the middle column of Table B.8. We attribute this to the temporal information insufficiency. Because the default frame rate of the used thermal camera is 90 fps, there is almost no movement between two adjacent frames. Therefore, the extracted temporal information is empty and

thus leading to an inability to distinguish standing from walking.

Table B.8: Classification accuracy on the testing set.

Partition	Accuracy (90 fps)(%)	Accuracy (30 fps)(%)
1	62.56	95.57
2	85.50	93.70
3	86.89	99.18
4	66.38	94.47
5	77.44	93.49
6	82.39	94.58
Average	76.86	95.17

To fix this problem, for every video, we sample one frame from every three frames and get the video of 30 fps. We redo the experiment with the same configurations as the 90 fps version experiment, achieving a significant improvement in the classification rate, see the last column in Table B.8.

For better insights into the influence of the frame rate on the activity recognition, the confusion matrices on partition 1 where the largest improvement happens are given in Fig. B.5. The upper matrix and the lower matrix correspond to 90 fps and 30 fps, respectively.

As there is no other work doing simultaneous recognition of clothes type and activity, we compare our results with works on clothes type recognition or activity recognition separately. FashionNet got an 82.58% top-3 classification accuracy on the DeepFashion dataset, an RGB garment dataset with 50 fine-grained categories [54]. A research group in Japan collected a thermal dataset including 5 categories (no action, walking, sitting down, standing up, and falling down) and gave recognition rates of 91.07% in 2017 [55] and 96.98% in 2020 [56]. Compared to these works, our method identifying both types with an average accuracy of 95.17% is effective.

4.3 Evaluation of Temperature Acquisition

The acquisition of the skin temperature and the clothes temperature critically relies on the body parts localization by detected body key points from OpenPose [10]. We evaluate the temperature acquisition module by examining the performance of OpenPose. As far as we know, this is the first evaluation of the assessment of applying OpenPose to thermal data quantitatively.

We randomly sample 10 short videos in each category and examine the performance of OpenPose on these videos. Specifically, as OpenPose is a frame-by-frame body key part detection algorithm, we save each processed frame and then count the number of unsatisfactory frames. It is worth mentioning that: (i) each frame is examined conservatively, meaning that a frame

4. Experiments

StandTshirts	4	0	0	0	0	0	0	0	0	0	27	0	0	0	0
StandLong	0	31	0	5	0	0	0	0	0	0	35	0	4	0	0
StandLongUnzip	0	0	21	0	0	0	0	0	0	0	0	14	0	0	0
StandRoll	0	0	0	12	0	0	0	0	1	0	0	0	0	66	0
StandRollUnzip	0	0	0	0	9	0	0	0	0	0	0	0	0	0	24
SitTshirts	0	0	0	0	0	25	0	0	3	0	5	0	0	0	0
SitLong	0	0	0	0	0	0	62	0	6	0	0	2	0	2	0
SitLongUnzip	0	0	0	0	0	0	0	23	0	0	0	0	8	0	0
SitRoll	0	0	0	0	0	0	0	0	58	1	0	0	0	15	0
SitRollUnzip	0	0	0	0	0	0	0	0	0	25	0	0	0	0	6
WalkTshirts	0	0	0	0	0	0	0	0	0	0	12	0	0	4	0
WalkLong	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0
WalkLongUnzip	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
WalkRoll	0	0	0	0	0	0	0	0	0	0	0	0	0	35	0
WalkRollUnzip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
	StandTshirts	StandLong	StandLongUnzip	StandRoll	StandRollUnzip	SitTshirts	SitLong	SitLongUnzip	SitRoll	SitRollUnzip	WalkTshirts	WalkLong	WalkLongUnzip	WalkRoll	WalkRollUnzip

(a)

StandTshirts	23	0	0	0	0	1	0	0	0	0	7	0	0	0	0
StandLong	0	73	0	2	0	0	0	0	0	0	0	0	0	0	0
StandLongUnzip	0	0	35	0	0	0	0	0	0	0	0	0	0	0	0
StandRoll	0	0	0	79	0	0	0	0	0	0	0	0	0	0	0
StandRollUnzip	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0
SitTshirts	0	0	0	0	0	31	0	0	2	0	0	0	0	0	0
SitLong	0	3	0	0	0	0	63	0	6	0	0	0	0	0	0
SitLongUnzip	0	0	4	0	0	0	0	27	0	0	0	0	0	0	0
SitRoll	0	0	0	0	0	0	0	0	72	2	0	0	0	0	0
SitRollUnzip	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0
WalkTshirts	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0
WalkLong	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0
WalkLongUnzip	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
WalkRoll	0	0	0	0	0	0	0	0	0	0	0	0	0	35	0
WalkRollUnzip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
	StandTshirts	StandLong	StandLongUnzip	StandRoll	StandRollUnzip	SitTshirts	SitLong	SitLongUnzip	SitRoll	SitRollUnzip	WalkTshirts	WalkLong	WalkLongUnzip	WalkRoll	WalkRollUnzip

(b)

Fig. B.5: Confusion matrices on partition 1. (a) 90 fps. (b) 30 fps.

with only one missing or wrongly detected body part is counted as an unsatisfactory frame; (ii) only the parts of the upper body (eyes, nose, neck, shoulders, elbows, and wrists) are examined as we aim at upper body parts for the reason that lower body parts are usually occluded by desks thus invisible.

Table B.9 lists the number of unsatisfactory frames (before “/”) and the number of all sampled frames (after “/”) within a certain category. The total number 805/14928 (0.0539) indicates that OpenPose can accurately detect human key body points on 94.61% of the sampled thermal frames, showing the reliability of applying OpenPose to the thermal dataset.

Table B.9: OpenPose performance on sampled frames.

-	Stand	Sit	Walk	Total
Long sleeves	8/1070	0/900	47/1004	55/2974
Long, unzipped	75/968	0/1070	99/888	174/2926
Rolled-up sleeves	89/917	69/1070	14/829	172/2816
Rolled-up, unzipped	0/1070	103/1070	128/1009	231/3149
T-shirt	107/973	25/1014	41/1076	173/3063
Total	279/4998	197/5124	329/4806	805/14928

Among the unsatisfactory frames, many of them belong to a few particular videos. For instance, 87 frames are related to wrongly detected eyes because the person bows his/her head (in Fig. B.6(a)). 80 frames are related to wrongly detected right wrist because the person overlaps his/her hands (in Fig. B.6(b)). Fortunately, these two types of wrongly detected body parts have minor influences on the skin region and clothes region segmentation described in Section 3.2.

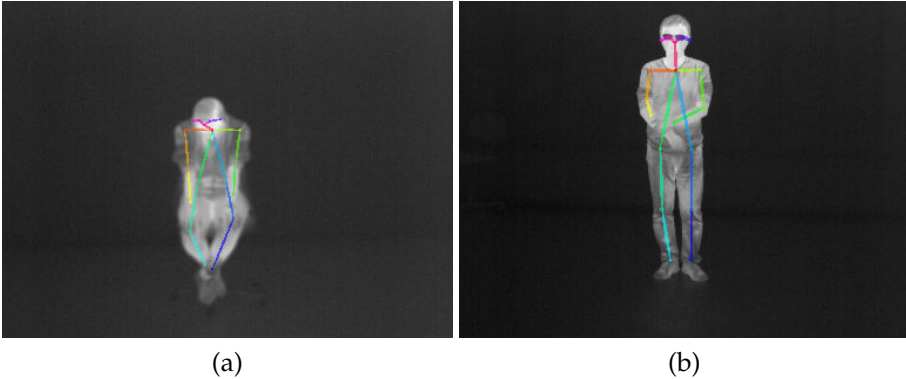


Fig. B.6: Two types of wrongly detected body parts. (a) Wrongly detected eyes. (b) Wrongly detected right wrist.

4. Experiments

As this is the first work evaluating OpenPose on thermal data quantitatively, we can only refer to the performance of OpenPose on the RGB mode as a comparison. For MPII dataset [57], an RGB dataset including about 25,000 images containing over 40,000 people, OpenPose achieves an average accuracy of 80.83% in detecting the upper body parts. Even though the RGB mode gives more visual details, the MPII dataset is extracted from YouTube videos and covers 410 activities like doing sports, playing music, taking medication, and so on. Therefore, it is more difficult than our thermal dataset but still indicates that our performance of 94.61% is reasonable and reliable.

4.4 Evaluation of I_{cl} and M Estimation

To assess the effectiveness of estimating I_{cl} and M from extracted key factors, we select a subject who wore all the five types of clothes during the dataset collection phase as a representative and list the calculated I_{cl} (in unit *clo*) and M (in unit W/m^2) in Table B.10.

The basic I_{cl} in the table is an average value over all the frames from a randomly selected short video, and the value in the round brackets is the variance. The final I_{cl} is the corrected value from the basic one.

From Table B.10, the same clothes usually have similar basic I_{cl} values despite different activities, revealing the intrinsic property of the clothes. Corrections considering activities or decreased clothing-covered regions introduce additional fluctuations in I_{cl} values. At the same time, the M values can be estimated directly from the occupant's activity type mentioned in Section 3.3 without extra equipment or human annotations.

Our estimations of I_{cl} and M are based on a contactless method without interference with subjects, unlike other related researches involving questionnaires, skin-attached sensors, weighing clothes, or wearable devices that cannot be applied to daily life. Therefore, to verify the feasibility of using our method in a real environment, we use CBE thermal comfort tool [58, 59] to compute the assessed human thermal comfort sensations of the subject, see the last four columns of Table B.10. This calculation is based on the PMV model mentioned in Section 1, and the other four environmental factors used in the calculation are air temperature (24°C), mean radiant temperature (24°C), air speed (0.1 m/s), and relative humidity (40%).

For better explanation, Table B.11 gives the meanings of PMV1 to PMV4. Specifically, PMV1 means the assessed thermal sensations when using fixed I_{cl} and M as 0.5 (a representative I_{cl} value in summer) and 55 (the M of sitting, the predominant activity in an office), respectively. PMV2 means the assessed thermal sensations when calculating I_{cl} and using a fixed M as 55. PMV3 means the assessed thermal sensations when using a fixed I_{cl} as 0.5 and calculating M . PMV4 means the assessed thermal sensations when calculating both I_{cl} and M . It is clear that by using the proposed estimation of

Table B.10: Estimated I_{cl} and M of one subject.

Clothes	Activity	Basic I_{cl}	Correction	Final I_{cl}	M	PMV1	PMV2	PMV3	PMV4
Long	Sitting	0.6361 (0.0138)	3	0.6648	55	-1.15	-0.73	-1.15	-0.73
Long	Standing	0.6830 (0.0092)	-	0.6830	70	-1.15	-0.68	-0.46	-0.08
Long	Walking	0.6999 (0.0559)	4	0.4694	150	-1.15	-1.23	1.03	0.98
Long, unzipped	Sitting	0.6441 (0.0072)	2,3	0.5987	55	-1.15	-0.88	-1.15	-0.88
Long, unzipped	Standing	0.5740 (0.0085)	2	0.4909	70	-1.15	-1.17	-0.46	-0.48
Long, unzipped	Walking	0.6266 (0.0190)	2,4	0.3412	150	-1.15	-1.64	1.03	0.75
Rolled-up	Sitting	0.6117 (0.0103)	1,3	0.5956	55	-1.15	-0.89	-1.15	-0.89
Rolled-up	Standing	0.6546 (0.0123)	1	0.6004	70	-1.15	-0.88	-0.46	-0.24
Rolled-up	Walking	0.7137 (0.0356)	1,4	0.4455	150	-1.15	-1.30	1.03	0.95
Rolled-up, unzipped	Sitting	0.4912 (0.0071)	1,2,3	0.4164	55	-1.15	-1.39	-1.15	-1.39
Rolled-up, unzipped	Standing	0.5305 (0.0078)	1,2	0.3932	70	-1.15	-1.46	-0.46	-0.73
Rolled-up, unzipped	Walking	0.5953 (0.0513)	1,2,4	0.2489	150	-1.15	-2.02	1.03	0.54
T-shirt	Sitting	0.1648 (0.0060)	3	0.2500	55	-1.15	-2.02	-1.15	-2.02
T-shirt	Standing	0.1657 (0.0075)	-	0.1657	70	-1.15	-2.43	-0.46	-1.47
T-shirt	Walking	0.1832 (0.0145)	4	0.0498	150	-1.15	-3.00	1.03	-0.16

4. Experiments

I_{cl} and M , the subject's assessed thermal sensations are much more dynamic which is more realistic and reasonable in daily life.

Table B.11: Meanings of PMV1 to PMV4.

-	Fixed I_{cl}	Proposed I_{cl} estimation
Fixed M	PMV1	PMV2
Proposed M estimation	PMV3	PMV4

Moreover, as in practice many engineers [17, 60–63] directly use I_{cl} values of typical garments (t-shirt, shirt, sleeveless vest, etc.) from ISO 8996 or ASHRAE 55 for simplicity, to have a further investigation and comparison, we also refer such I_{cl} values from ASHRAE 55 as basic reference I_{cl} in Table B.12, where 0.5600 refers to two layers of clothes (a t-shirt and a winter coat) and 0.0800 refers to a t-shirt. Also in Table B.12, basic I_{cl} , final I_{cl} , and PMV4 are the same as those in Table B.10; final reference I_{cl} represents the corrected values from basic reference I_{cl} based on the proposed correction 1 to 4; PMV5 refers to the assessed thermal sensations when using the basic reference I_{cl} and calculating M ; PMV6 refers to the assessed thermal sensations when using the final reference I_{cl} and calculating M . From Table B.12, the difference between reference I_{cl} and estimated I_{cl} is usually less than 0.1 which is equivalently the insulation rate of a t-shirt, well showing the consistency and acceptance of the proposed method. Besides, when considering the influences of rolled-up sleeves, unzipped zippers, and different activities, the final reference I_{cl} is more dynamic, which makes it more applicable to real indoor environments.

To best illustrate the comparison of assessed thermal sensations (PMV1 to PMV6), Fig. B.7 draws the sensations as colorful circles in a more readable way based on the standard 7 scales. It is obvious to see that with fixed I_{cl} and M (see Fig. B.7(a)), the subject is always regarded to be slightly cool, suggesting that the indoor microclimate should be warmer. A similar phenomenon happens when M is fixed in PMV2 (see Fig. B.7(b)). When only fixing I_{cl} , resultant sensations in PMV3 (see Fig. B.7(c)) seems plausible; however, feeling slightly warm when walking with a short-sleeved t-shirt in a 24°C room is contradicting the real sensation stated by the subject when collecting the dataset. After calculating both I_{cl} and M , PMV4 (see Fig. B.7(d)) is much more reasonable, as the thermal sensations change dynamically from cool to slightly warm with the increase of the clothes' thermal insulation ability and the activity intensity. When taking basic reference I_{cl} into consideration, PMV5 (see Fig. B.7(e)) is more acceptable than PMV3, but it is still difficult to describe the real situation as the basic reference I_{cl} ignores the different status of sleeves and zippers. As a contrast, corrected reference I_{cl} in PMV6 (see Fig. B.7(f)) makes assessed thermal feelings dynamic and realistic. By comparing

Table B.12: Comparison between reference I_{cl} and estimated I_{cl} of one subject.

Clothes	Activity	Basic I_{cl}	Basic reference I_{cl}	Final I_{cl}	Final reference I_{cl}	PMV4	PMV5	PMV6
Long	Sitting	0.6361	0.5600	0.6648	0.5978	-0.73	-0.98	-0.88
Long	Standing	0.6830	0.5600	0.6830	0.5600	-0.08	-0.33	-0.33
Long	Walking	0.6999	0.5600	0.4694	0.3607	0.98	1.12	0.79
Long, unzipped	Sitting	0.6441	0.5600	0.5987	0.5246	-0.88	-0.98	-1.08
Long, unzipped	Standing	0.5740	0.5600	0.4909	0.4769	-0.48	-0.33	-0.52
Long, unzipped	Walking	0.6266	0.5600	0.3412	0.2680	0.75	1.12	0.58
Rolled-up	Sitting	0.6117	0.5600	0.5956	0.5501	-0.89	-0.98	-1.01
Rolled-up	Standing	0.6546	0.5600	0.6004	0.5058	-0.24	-0.33	-0.45
Rolled-up	Walking	0.7137	0.5600	0.4455	0.2988	0.95	1.12	0.66
Rolled-up, unzipped	Sitting	0.4912	0.5600	0.4164	0.4770	-1.39	-0.98	-1.21
Rolled-up, unzipped	Standing	0.5305	0.5600	0.3932	0.4227	-0.73	-0.33	-0.65
Rolled-up, unzipped	Walking	0.5953	0.5600	0.2489	0.2149	0.54	1.12	0.44
T-shirt	Sitting	0.1648	0.0800	0.2500	0.1754	-2.02	-2.94	-2.38
T-shirt	Standing	0.1657	0.0800	0.1657	0.0800	-1.47	-1.86	-1.86
T-shirt	Walking	0.1832	0.0800	0.0498	0.0136	-0.16	-0.03	-0.35

5. Conclusions and Future Work

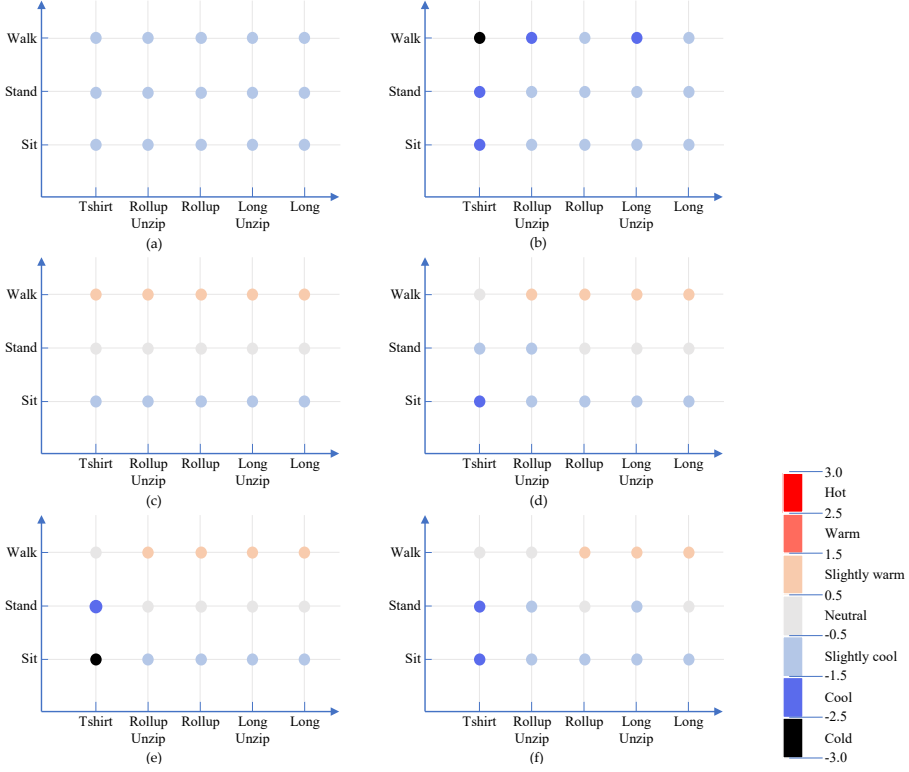


Fig. B.7: Assessed thermal comfort sensations. (a) PMV1. (b) PMV2. (c) PMV3. (d) PMV4. (e) PMV5. (f) PMV6.

PMV4 (Fig. B.7(d)) and PMV6 (Fig. B.7(f)), 12/15 (80%) feelings are same, revealing the consistency between ISO-based I_{cl} and our proposed estimation method. As a whole, these comparisons emphasize the practicability of our method in a daily environment.

5 Conclusions and Future Work

In this paper, an occupant's clothes type, activity type, skin temperature, and clothes temperature are extracted by vision-based procedures. These personal factors are used to automatically estimate individual clothing insulation rate I_{cl} and metabolic rate M , which are critical for dynamic thermal comfort assessment. Specifically, with the captured thermal videos as inputs, a convolutional neural network (CNN) is implemented to recognize an occupant's clothes type and activity type simultaneously. With key body parts detected by OpenPose as localization references, the recognized clothes type can help

to separate a human body into the skin-bare regions and the clothing-covered regions that correspond to the skin temperature and the clothes temperature. From these extracted personal factors, a comprehensive estimation of I_{cl} and M according to the international standards is implemented. Experiments evaluating the CNN module, the temperature calculation module, and the I_{cl}/M estimation module are also given. By our method, the assessed thermal sensations are dynamically changing according to the estimated I_{cl} and M , which is reasonable and realistic in daily life, well proving the feasibility of the proposed scheme.

Future work includes taking additional clothes types, activity types, and other physiological aspects into consideration and expanding the proposed scheme in field studies of multi-person for better microclimate control.

6 Appendix

According to Annex H of ISO 9920 [18], I_{cl} can be calculated from the mass of the clothes and the body surface area covered by clothing, that is

$$I_{cl} = 0.919 + 0.255 \times 10^{-3} \cdot m - 0.00874 \cdot A_{cov,0} - 0.00510 \cdot A_{cov,1} \quad (B.9)$$

where, m is the mass of the clothes, without shoes, in grams; $A_{cov,0}$ is the body surface area not covered by clothing; $A_{cov,1}$ is the body surface area covered by a single clothing layer; both $A_{cov,0}$ and $A_{cov,1}$ are expressed as percentages of the total body surface area shown in Table B.13.

When a person rolls up the sleeves or unzips the zipper, the mass of the clothes remains the same, while the $A_{cov,0}$ or the $A_{cov,1}$ is changed, leading to a correction in I_{cl} by Equations B.4, B.5 or B.6 in Section 3.3.

According to Section 9 of ISO 9920, when a person is sitting, the compressed air between the body and the clothes leads to a decrease in I_{cl} by 6% to 18%. Office chairs introduce an increase in I_{cl} of 0.04 *clo* to 0.17 *clo*. We use the median values of these fluctuations as 12% and 0.105 *clo*, respectively, and get Equation B.7 in Section 3.3.

According to Section 8.2 of ISO 9920, the dynamic clothing insulation rate $I_{cl,d}$ under a condition having air and body movement can be calculated from

$$I_{cl,d} = \frac{(0.6 - I_{cl,s}) \cdot I_{a,d} + I_{cl,s} \cdot I_{t,d}}{0.6} - I_{a,d} \quad (B.10)$$

$$(0 < I_{cl,s} \leq 0.6)$$

$$I_{cl,d} = I_{t,d} - I_{a,d} (0.6 < I_{cl,s} < 1.4) \quad (B.11)$$

$$I_{t,d} = I_{t,s} \cdot C_t \quad (B.12)$$

6. Appendix

Table B.13: Body surface area percentage [18].

Segment	% total area
Head + neck	8.7
Chest	10.2
Back	9.2
Abdomen	6.1
Buttocks	6.6
Right upper arm	5.0
Left upper arm	5.0
Right lower arm	3.1
Left lower arm	3.1
Right hand	2.5
Left hand	2.5
Right thigh	9.2
Left thigh	9.2
Right calf	6.1
Left calf	6.1
Right foot	3.7
Left foot	3.7
Total	100

$$I_{a,d} = I_{a,s} \cdot C_a \quad (\text{B.13})$$

$$I_{t,s} = I_{cl,s} + I_{a,s} \quad (\text{B.14})$$

$$C_t = e^{[-0.281 \cdot (v_a - 0.15) + 0.044 \cdot (v_a - 0.15)^2 - 0.492 \cdot v_w + 0.176 \cdot v_w^2]} \quad (\text{B.15})$$

$$C_a = e^{[-0.533 \cdot (v_a - 0.15) + 0.069 \cdot (v_a - 0.15)^2 - 0.462 \cdot v_w + 0.201 \cdot v_w^2]} \quad (\text{B.16})$$

where, $I_{t,d}$, $I_{a,d}$, $I_{cl,s}$, $I_{t,s}$, $I_{a,s}$, v_a , and v_w are the dynamic total thermal insulation, the dynamic air insulation, the static clothes insulation, the static total insulation, the static air insulation, the relative air velocity in relation to human motion, and the human motion velocity, respectively.

As Section 6 of ISO 9920 mentions, the static air insulation $I_{a,s}$ in most studies is around 0.7 *clo*. When the human motion velocity is 1 m/s, for a wind-free environment (air velocity is smaller than 0.2 m/s), the relative air velocity in relation to human motion is about 1 m/s, and then from Equations B.10-B.16, Equation B.8 in Section 3.3 is derived.

References

- [1] US EIA, “Residential energy consumption survey,” *US Energy Information Administration* 2009, 2005.
- [2] Povl Ove Fanger, “Assessment of man’s thermal comfort in practice,” *Occupational and Environmental Medicine*, vol. 30, no. 4, pp. 313–324, 1973.
- [3] International Standard Organization, “Ergonomics of the thermal environment — analytical determination and interpretation of thermal comfort using calculation of the pmv and ppd indices and local thermal comfort criteria,” <https://www.iso.org/standard/39155.html/>, last accessed: May, 2020.
- [4] Jian Liang and Ruxu Du, “Thermal comfort control based on neural network for hvac application,” in *Proceedings of 2005 IEEE Conference on Control Applications*, 2005. CCA 2005. IEEE, 2005, pp. 819–824.
- [5] Ruey-Lung Hwang and Shiu-Ya Shu, “Building envelope regulations on thermal comfort in glass facade buildings and energy-saving potential for pmv-based comfort control,” *Building and Environment*, vol. 46, no. 4, pp. 824–834, 2011.
- [6] Kenta Kuzuhara and Hiroaki Nishi, “Accurate indoor condition control based on pmv prediction in bems environments,” in *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2013, pp. 8142–8147.
- [7] Eusébio ZE Conceição, João MM Gomes, and António E Ruano, “Application of hvac systems with control based on pmv index in university buildings with complex topology,” *IFAC-PapersOnLine*, vol. 51, no. 10, pp. 20–25, 2018.
- [8] Eusébio ZE Conceição, António FM Sousa, João MM Gomes, and António E Ruano, “Hvac systems applied in university buildings with control based on pmv and apmv indexes,” *Inventions*, vol. 4, no. 1, pp. 3, 2019.
- [9] Jing Wu, Xiangdong Li, Jiyuan Tu, Lin Yang, and Yihuan Yan, “A pmv-based hvac control strategy for office rooms subjected to solar radiation,” *Building and Environment*, p. 106863, 2020.
- [10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

- [11] ASHRAE, "Standard55-thermal environmental conditions for human occupancy," http://arco-hvac.ir/wp-content/uploads/2015/11/ASHRAE_Thermal_Comfort_Standard.pdf, last accessed: May, 2020.
- [12] Jack Ngarambe, Geun Young Yun, and Gon Kim, "Prediction of indoor clothing insulation levels: A deep learning approach," *Energy and Buildings*, vol. 202, pp. 109402, 2019.
- [13] Fergus Nicol and Susan Roaf, "Pioneering new indoor temperature standards: the pakistan project," *Energy and Buildings*, vol. 23, no. 3, pp. 169–174, 1996.
- [14] Michele De Carli, Bjarne W Olesen, Angelo Zarrella, and Roberto Zecchin, "People's clothing behaviour according to external weather and indoor environment," *Building and Environment*, vol. 42, no. 12, pp. 3965–3973, 2007.
- [15] Frédéric Haldi and Darren Robinson, "Modelling occupants' personal characteristics for thermal comfort prediction," *International journal of biometeorology*, vol. 55, no. 5, pp. 681–694, 2011.
- [16] Paulo Matos de Carvalho, Manuel Gameiro da Silva, and João Esteves Ramos, "Influence of weather and indoor climate on clothing of occupants in naturally ventilated school buildings," *Building and environment*, vol. 59, pp. 38–46, 2013.
- [17] Weiwei Liu, Diyu Yang, Xiong Shen, and Peizhi Yang, "Indoor clothing insulation and thermal history: a clothing model based on logistic function and running mean outdoor temperature," *Building and Environment*, vol. 135, pp. 142–152, 2018.
- [18] International Standard Organization, "Ergonomics of the thermal environment — estimation of thermal insulation and water vapour resistance of a clothing ensemble," <https://www.iso.org/standard/39257.html/>, last accessed: May, 2020.
- [19] Maria Konarska, Krzysztof Soltynski, Iwona Sudol-Szopinska, and Anna Chojnacka, "Comparative evaluation of clothing thermal insulation measured on a thermal manikin and on volunteers," *Fibres and Textiles in Eastern Europe*, vol. 15, no. 2, pp. 73, 2007.
- [20] Siliang Lu and Erica Cochran Hameen, "Integrated ir vision sensor for online clothing insulation measurement," 2018.
- [21] BW Olesen and R Nielsen, "Thermal insulation of clothing measured on a movable thermal manikin and on human subjects," *ECSC Programme Research*, , no. 7206/00, pp. 914, 1983.

- [22] Elizabeth A McCullough, Byron W Jones, and Janice Huck, "A comprehensive data base for estimating clothing insulation," *Ashrae Trans*, vol. 91, no. 2, pp. 29–47, 1985.
- [23] Hiroki Matsumoto, Yoshio Iwai, and Hiroshi Ishiguro, "Estimation of thermal comfort by measuring clo value without contact.," in *MVA*. Citeseer, 2011, pp. 491–494.
- [24] Jeong-Hoon Lee, Young-Keun Kim, Kyung-Soo Kim, and Soohyun Kim, "Estimating clothing thermal insulation using an infrared camera," *Sensors*, vol. 16, no. 3, pp. 341, 2016.
- [25] Jun Miura, Mitsuhiro Demura, Kaichiro Nishi, and Shuji Oishi, "Thermal comfort measurement using thermal-depth images for robotic monitoring," *Pattern Recognition Letters*, 2019.
- [26] Kyungsoo Lee, Haneul Choi, Hyungkeun Kim, Daeung Danny Kim, and Taeyeon Kim, "Assessment of a real-time prediction method for high clothing thermal insulation using a thermoregulation model and an infrared camera," *Atmosphere*, vol. 11, no. 1, pp. 106, 2020.
- [27] Maohui Luo, Xiang Zhou, Yingxin Zhu, and Jan Sundell, "Revisiting an overlooked parameter in thermal comfort studies, the metabolic rate," *Energy and Buildings*, vol. 118, pp. 152–159, 2016.
- [28] Yongchao Zhai, Minghui Li, Siru Gao, Liu Yang, Hui Zhang, Edward Arens, and Yunfei Gao, "Indirect calorimetry on the metabolic rate of sitting, standing and walking office activities," *Building and Environment*, vol. 145, pp. 77–84, 2018.
- [29] Wenjie Ji, Maohui Luo, Bin Cao, Yingxin Zhu, Yang Geng, and Borong Lin, "A new method to study human metabolic rate changes and thermal comfort in physical exercise by co2 measurement in an airtight chamber," *Energy and Buildings*, vol. 177, pp. 402–412, 2018.
- [30] Andrea Calvaresi, Marco Arnesano, Filippo Pietroni, and Gian Marco Revel, "Measuring metabolic rate to improve comfort management in buildings.," *Environmental Engineering & Management Journal (EEMJ)*, vol. 17, no. 10, 2018.
- [31] Sara Casaccia, Filippo Pietroni, Andrea Calvaresi, Gian Marco Revel, and Lorenzo Scalise, "Smart monitoring of user's health at home: Performance evaluation and signal processing of a wearable sensor for the measurement of heart rate and breathing rate.," in *BIOSIGNALS*, 2016, pp. 175–182.

- [32] HooSeung Na, Joon-Ho Choi, HoSeong Kim, and Taeyeon Kim, "Development of a human metabolic rate prediction model based on the use of kinect-camera generated visual data-driven approaches," *Building and Environment*, vol. 160, pp. 106216, 2019.
- [33] Mohammad H Hasan, Fadi Alsaleem, and Mostafa Rafaie, "Sensitivity study for the pmv thermal comfort model and the use of wearable devices biometric data for metabolic rate estimation," *Building and Environment*, vol. 110, pp. 173–183, 2016.
- [34] Syed Ihtsham-ul-Haq Gilani, Muhammad Hammad Khan, and Muzaffar Ali, "Revisiting fanger's thermal comfort model using mean blood pressure as a bio-marker: An experimental investigation," *Applied Thermal Engineering*, vol. 109, pp. 35–43, 2016.
- [35] International Standard Organization, "Ergonomics of the thermal environment — determination of metabolic rate," <https://www.iso.org/standard/34251.html/>, last accessed: May, 2020.
- [36] International Standard Organization, "Ergonomics of the thermal environment — analytical determination and interpretation of heat stress using calculation of the predicted heat strain," <https://www.iso.org/standard/37600.html/>, last accessed: May, 2020.
- [37] Xiaogang Cheng, Bin Yang, Thomas Olofsson, Guoqing Liu, and Haibo Li, "A pilot study of online non-invasive measuring technology based on video magnification to determine skin temperature," *Building and Environment*, vol. 121, pp. 1–10, 2017.
- [38] Xiaogang Cheng, Bin Yang, Anders Hedman, Thomas Olofsson, Haibo Li, and Luc Van Gool, "Non-invasive thermal comfort perception based on subtleness magnification and deep learning for energy efficiency," *arXiv preprint arXiv:1811.08006*, 2018.
- [39] Farrokh Jazizadeh and Wooyoung Jung, "Personalized thermal comfort inference using rgb video images for distributed hvac control," *Applied Energy*, vol. 220, pp. 829–841, 2018.
- [40] Da Li, Carol C Menassa, and Vineet R Kamat, "Feasibility of low-cost infrared thermal imaging to assess occupants' thermal comfort," in *Computing in Civil Engineering 2019: Smart Cities, Sustainability, and Resilience*, pp. 58–65. American Society of Civil Engineers Reston, VA, 2019.
- [41] Da Li, Carol C Menassa, and Vineet R Kamat, "Robust non-intrusive interpretation of occupant thermal comfort in built environments with low-cost networked thermal cameras," *Applied energy*, vol. 251, pp. 113336, 2019.

- [42] Siliang Lu, Weilong Wang, Shihan Wang, and Erica Cochran Hameen, "Thermal comfort-based personalized models with non-intrusive sensing technique in office buildings," *Applied Sciences*, vol. 9, no. 9, pp. 1768, 2019.
- [43] Junpeng Qian, Xiaogang Cheng, Bin Yang, Zhe Li, Junchi Ren, Thomas Olofsson, and Haibo Li, "Vision-based contactless pose estimation for human thermal discomfort," *Atmosphere*, vol. 11, no. 4, pp. 376, 2020.
- [44] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.
- [45] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 363–378.
- [46] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.
- [47] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, pp. 214–223.
- [48] Jinsong Liu, Isak Worre Foged, and Thomas B Moeslund, "Vision-based individual factors acquisition for thermal comfort assessment in a built environment," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 662–666.
- [49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [50] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [51] Changzhi Dai, Hui Zhang, Edward Arens, and Zhiwei Lian, "Machine learning approaches to predict thermal demands using skin temperatures: Steady-state conditions," *Building and Environment*, vol. 114, pp. 1–10, 2017.
- [52] Andrei Claudiu Cosma and Rahul Simha, "Thermal comfort modeling in transient conditions using real-time local body temperature extraction

- with a thermographic camera," *Building and Environment*, vol. 143, pp. 36–47, 2018.
- [53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [54] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deep-fashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.
- [55] Takayuki Kawashima, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase, Daisuke Deguchi, Tomoyoshi Aizawa, and Masato Kawade, "Action recognition from extremely low-resolution thermal image sequence," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.
- [56] Igor Morawski and Wen-Nung Lie, "Two-stream deep learning architecture for action recognition by using extremely low-resolution infrared thermopile arrays," in *International Workshop on Advanced Imaging Technology (IWAIT) 2020*. International Society for Optics and Photonics, 2020, vol. 11515, p. 115150Y.
- [57] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.
- [58] Tyler Hoyt, Stefano Schiavon, Federico Tartarini, Toby Cheung, Kyle Steinfeld, Alberto Piccioli, and Dustin Moon, "Cbe thermal comfort tool," <https://comfort.cbe.berkeley.edu/EN>, last accessed: November, 2020.
- [59] Federico Tartarini, Stefano Schiavon, Toby Cheung, and Tyler Hoyt, "Cbe thermal comfort tool: online tool for thermal comfort calculations and visualizations," *SoftwareX*, vol. 12, pp. 100563, 2020.
- [60] Yu Jiao, Hang Yu, Tian Wang, Yusong An, and Yifan Yu, "The relationship between thermal environments and clothing insulation for elderly individuals in shanghai, china," *Journal of thermal biology*, vol. 70, pp. 28–36, 2017.
- [61] Radostina A Angelova, Elena Georgieva, Detelin Markov, Tsvetan Bozhkov, Iskra Simova, Nushka Kehaiova, and Peter Stankov, "Estimating the effect of torso clothing insulation on body skin and clothing

References

- temperatures in a cold environment using infrared thermography," *Fibres & Textiles in Eastern Europe*, 2018.
- [62] Ferdinando Salata, Iacopo Golasi, Virgilio Ciancio, and Federica Rosso, "Dressed for the season: Clothing and outdoor thermal comfort in the mediterranean population," *Building And Environment*, vol. 146, pp. 50–63, 2018.
- [63] Tianyu Xi, Qiaochu Wang, Huan Qin, and Hong Jin, "Influence of outdoor thermal environment on clothing and activity of tourists and local people in a severely cold climate city," *Building and Environment*, vol. 173, pp. 106757, 2020.

Paper C

Clothing Insulation Rate and Metabolic Rate Estimation for Individual Thermal Comfort Assessment in Real Life

Jinsong Liu, Isak Worre Foged, and Thomas B. Moeslund

The paper has been published in the journal of
Sensors

© 2022 Authors

The layout has been revised.

Abstract

Satisfactory indoor thermal environments can improve working efficiencies of office staff. To build such satisfactory indoor microclimates, individual thermal comfort assessment is important, for which personal clothing insulation rate (I_{cl}) and metabolic rate (M) need to be estimated dynamically. Therefore, this paper proposes a vision-based method. Specifically, a human tracking-by-detection framework is implemented to acquire each person's clothing status (short-sleeved, long-sleeved), key posture (sitting, standing), and bounding box information simultaneously. The clothing status together with a key body points detector locate the person's skin region and clothes region, allowing the measurement of skin temperature (T_s) and clothes temperature (T_c), and realizing the calculation of I_{cl} from T_s and T_c . The key posture and the bounding box change across time can category the person's activity intensity into a corresponding level, from which the M value is estimated. Moreover, we have collected a multi-person thermal dataset to evaluate the method. The tracking-by-detection framework achieves a mAP_{50} (Mean Average Precision) rate of 89.1% and a MOTA (Multiple Object Tracking Accuracy) rate of 99.5%. The I_{cl} estimation module gets an accuracy of 96.2% in locating skin and clothes. The M estimation module obtains a classification rate of 95.6% in categorizing activity level. All of these prove the usefulness of the proposed method in a multi-person scenario of real-life applications.

keywords: thermal comfort; clothing insulation rate; metabolic rate; multi-person; real life

1 Introduction

In the world today, more people have to rely on computers to tackle various tasks. This results in indoor office work being much more popular than ever before. From the commercial buildings energy consumption survey in 2012 [1], offices consume much more energy for heating and cooling than other types of buildings. If energy can be used according to office workers' thermal needs, energy waste resulting from overheating or overcooling will be greatly reduced, and also staff will have better working efficiencies as they feel comfortable with the environment they work in.

To make each office staff feel thermal comfort and at the same time reduce energy waste, two main kinds of methods have been researched. One is directly relying on the worn clothes to control a person's micro-environment between the body skin and the indoor atmosphere, which avoids controlling the entire indoor microclimate via heaters, ventilation, and air conditioners (HVAC) that consume lots of energy. This kind of method takes advantage of different thermal properties (thermal resistance, thermal conductivity,

thermal radiation, thermal convection, water evaporation, etc.) of different clothes in materials, thicknesses, and layers to maintain the body temperature in a comfortable range [2–5]. The other kind of method still focuses on the entire indoor environment but in a way that adjusts the microclimate according to each occupant’s thermal need, which is the topic of this paper.

However, each person’s thermal need is unique and dynamic, which cannot be met well by existing microclimate-controlling systems like HVAC that all rely on static assumptions to serve the occupants in a room. For example, a standard air conditioner’s temperature is set to 25 to 27 degrees for cooling in summer, and 18 to 20 degrees for heating in winter, no matter whether this is what the office workers need.

To improve this situation, individual thermal comfort feeling has to be assessed, like in scales (cold, cool, slightly cool, neutral, slightly warm, warm, and hot) [6–8]. These scales depend on both environmental factors and personal factors. The environmental factors are air temperature (t_a), mean radiation temperature (\bar{t}_r), relative humidity (RH), and air velocity (V_a), which can be measured by sensors. The personal factors include clothing insulation rate (I_{cl}) and metabolic rate (M); I_{cl} describes the ability of the clothes to insulate the heat exchange between the skin and the environment outside the clothes, and M describes the amount of energy, in-unit time, consumed by a person. Both the personal factors are difficult to acquire for their complexity and dynamics.

Accordingly, international standards [8–12] have defined reference values of I_{cl} and M in certain situations (see Tables C.1 and C.2). Such values are empirical and fixed, and thus cannot describe a person’s dynamic property for that the situation in real life is much more complex than these noted ones. This hinders the development of systems and applications for adjusting indoor microclimates according to occupants’ thermal needs. Therefore, the solution dynamically estimating a person’s I_{cl} and M is to be explored. To this end, we propose a method to do this, and the concrete contributions are:

- The method inventively adapts state-of-the-art computer vision solutions to the thermal comfort domain, achieving a contactless approach that can be employed in multi-person real-life applications.
- The method can detect and track each person, at the same time recognizing his or her clothing status (long-sleeved, short-sleeved) and key posture (sitting, standing).
- The method can further output a person’s skin temperature and clothes temperature, based on which his or her I_{cl} is estimated.
- The method proposes three useful features from a person’s bounding box tracked across time. These features can category the person’s activity into a certain intensity level which indicates the M .

2. Related Work

Table C.1: Insulation values of various typical garments [10].

Garment		I_{cl} (clo)
Underwear	Singlet	0.04
	T-shirt	0.09
	Shirts with long sleeves	0.12
Shirts, blouses	Short sleeves	0.15
	Lightweight, long sleeves	0.2
	Normal, long sleeves	0.25

Table C.2: Metabolic rates of typical activities [8].

Activity	M (W/m ²)
Reclining	46
Seated, relaxed	58
Sedentary activity	70
Standing, light activity	93
Standing, medium activity	116
Walking on level ground:	
2 km/h	110
3 km/h	140
4 km/h	165
5 km/h	200

The rest contents are organized as follows. Section 2 introduces the related work. Section 3 describes our methodology. Section 4 tells the experiments. Section 5 concludes the paper and proposes future work.

2 Related Work

This paper applies computer vision solutions to the thermal comfort domain. Therefore, the related researches of both I_{cl} and M estimation and computer vision methods are studied.

2.1 I_{cl} and M Estimation

Several works have been published to calculate the two personal factors, I_{cl} and M , for assessing the human thermal sensation. However, most works only focus on one of them, leaving the other one unsolved.

Some works take advantage of the relationship between clothing choice and environment temperature [13–16] to predict clothing insulation ability. This type of method is simple but neglects the inherent property of clothes

themselves. To resolve this drawback, work [17] uses the weight of the clothes to predict I_{cl} , which is unrealistic in real applications; studies estimate I_{cl} from the temperature difference between the body skin and the clothes surface with infrared sensors [18, 19], however, this is also inconvenient due to the attached sensors on the human body. To decouple such interference with personal life, researches [20–22] all adopt contactless infrared cameras to monitor persons. Unfortunately, refs. [20, 21] do not mention the method of acquiring temperatures of interested body locations, limiting their applications in the real world; ref. [22] only considers five types of garments that cannot represent various clothing choices in daily life.

For metabolic rate estimation, almost all works have to use attached equipment. Correspondingly, a person’s M is estimated by measuring his or her oxygen consumption and carbon dioxide generation [23–25], heart rates [26–29], or blood pressure [30]. Though [31–33] adopt cameras for such a task, they still partly rely on sophisticated equipment mentioned above. These devices have to be worn by subjects, making them unrealistically used in daily life.

When estimating both I_{cl} and M , refs. [34, 35] use a CNN (Convolutional Neural Network)-based classifier to recognize a person’s clothes type and activity type, and then refer ISO (International Organization for Standardization) standard tables to get the I_{cl} and M values from the recognized types. These works prove the importance of clothing status (short sleeves, long sleeves) and posture (sitting, standing) in estimating I_{cl} and M . However, refs. [34, 35] are only valid in a simple and controlled single-person environment. Expanding and enriching this kind of solution is in great need. Therefore, this paper closes this gap and is the first work targeted at a multi-person scenario in the real world.

2.2 Detection and Tracking

The ability to do individual processing from multiple persons is the crucial point of the proposed method, which mainly comes from our implemented human tracking-by-detection framework. To this end, widely used object detectors are studied, like Faster R-CNN (Region-based Convolutional Neural Network) [36], YOLO (You Only Look Once) series [37–41], and FPN (Feature Pyramid Network) [42] which all consist of a backbone network (to extract deep features) and headers (to predict bounding box locations and categories). All these methods perform well on RGB (Red Green Blue) benchmark datasets [43, 44].

When it comes to the tracking part (referring in particular to online multi-object tracking in this paper), SORT (Simple Online and Realtime Tracking) [45] initially replaces the conventional object detector with a CNN-based detector and thus improves the tracking result by up to 18.9%, revealing the

importance of accurate detections for tracking. The following DeepSort (Simple Online and Realtime Tracking with a Deep Association Metric) [46] and CDA_DDAL (Confidence-based Data Association and Discriminative Deep Appearance Learning) [47] incorporate appearance information into the data association phase and solve the ID (Identity)-switch problem. Other works focus on improving the correlation filter to estimate better positions of targets in the next frame [48], fusing multi-modality data in data association [49], and linking detection and tracking to let them benefit each other [50].

In general, though existing methods on human detection and tracking are quite mature in RGB datasets, studies applying them in thermal datasets like [51–53] are few and far between. This situation makes our research with the thermal camera more essential.

3 Methodology

In this section, we describe our approach, the overview of which is illustrated in Fig. C.1 including three key parts:

1. The thermal input goes through a tracking-by-detection framework (see the red dashed box) to track each individual (see the ID 1 and ID 2) and at the same time categorize each person to get his or her clothing status and key posture (see the red and green solid boxes around persons which indicate different categories).
2. With ID information, for each person, the clothing status classified by the tracking-by-detection part helps differentiate the skin region from the clothing-covered region. Then the detected key body points from these two regions can represent the skin temperature and the clothes temperature, based on which I_{cl} is estimated.
3. With ID information, for each person, the optical flow within each person's bounding box region, together with the bounding box (center location and box size) changes across time are calculated. These three features are good representations of the person's activity intensity, which are used to estimate M .

Details of the three parts are described below.

3.1 Tracking-by-Detection

This part has two main components, one is an object detector, YOLOv5 [41], for human detection, the other is a tracker, DeepSort [46].

The video collected from a thermal camera is the input to the detector YOLOv5 for frame-by-frame human detection. To integrate clothing status

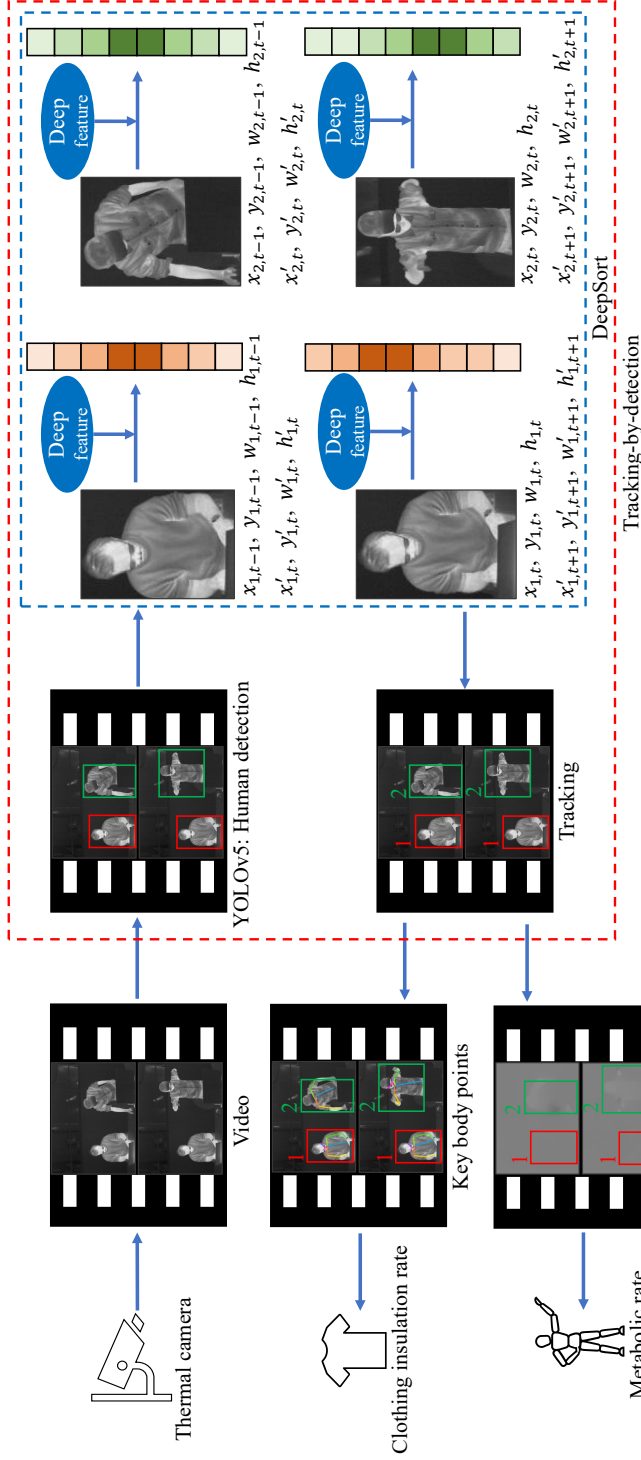


Fig. C.1: Overview of the proposed method. The numbers 1 and 2 are the corresponding tracking ID numbers of the two persons.

3. Methodology

and key posture recognition into this detection procedure, we classify persons into six categories (see Table C.3). Here the clothing status is represented by the sleeve status (long, short) for four reasons: (i) these two are the most common clothing situations in an office environment while the lower part of the body is often totally occluded by the desk; (ii) according to [10, 34, 35], sleeve status is significantly important in estimating I_{cl} ; (iii) the change between a long-sleeved status to a short-sleeved status by rolling up sleeves or taking off outer jackets is a sign of feeling hot and vice versa, indicating a person's thermal sensation directly; (iv) the sleeves status helps to locate skin region and clothes region separately for further skin and clothes temperatures acquisition. For example, the elbows of a person wearing short-sleeved clothes are skin regions, while the elbows of a person wearing long-sleeved clothes are clothes regions. This localization makes it possible to use such key body points to calculate a person's skin temperature and clothes temperature, because key body points on arms are widely used sensitive heat receptors in thermal comfort assessment [35, 54–56]. Besides the two statuses of long sleeves and short sleeves, another status called difficult to predict clothes type due to occlusion is also usual in daily life. For clear illustration, such cases are in Fig. C.2. The right persons in Fig. C.2(a) and Fig. C.2(b) are partly occluded by the computer monitor; the right person in Fig. C.2(c) moves the arms out of the scene; the left person in Fig. C.2(d) occludes his lower arms by hiding them behind the torso. These occlusions make it unrealistic to know whether the sleeves are long or short. One thing to be noted is that even though a person is occluded in a few frames, his or her clothing status can be recognized in other frames. Therefore, voting of a classified category over a few seconds is important. When it comes to the key posture recognition, from ISO standards [8, 9, 11, 12], a person's metabolic rate M is closely related to the behaving posture (sitting, standing, lying down, etc.). And in a typical office environment the most common ones are sitting and standing, therefore, these two are considered in our study.

Table C.3: Persons in six categories.

Category	Meaning
LongSit	Long-sleeved clothes, sitting
ShortSit	Short-sleeved clothes, sitting
OclSit	Difficult to predict clothes type due to occlusion, sitting
LongStand	Long-sleeved clothes, standing
ShortStand	Short-sleeved clothes, standing
OclStand	Difficult to predict clothes type due to occlusion, standing

The ultimate goal of this research is to acquire every occupant's personal factors and thus facilitate individual thermal comfort assessment. This means

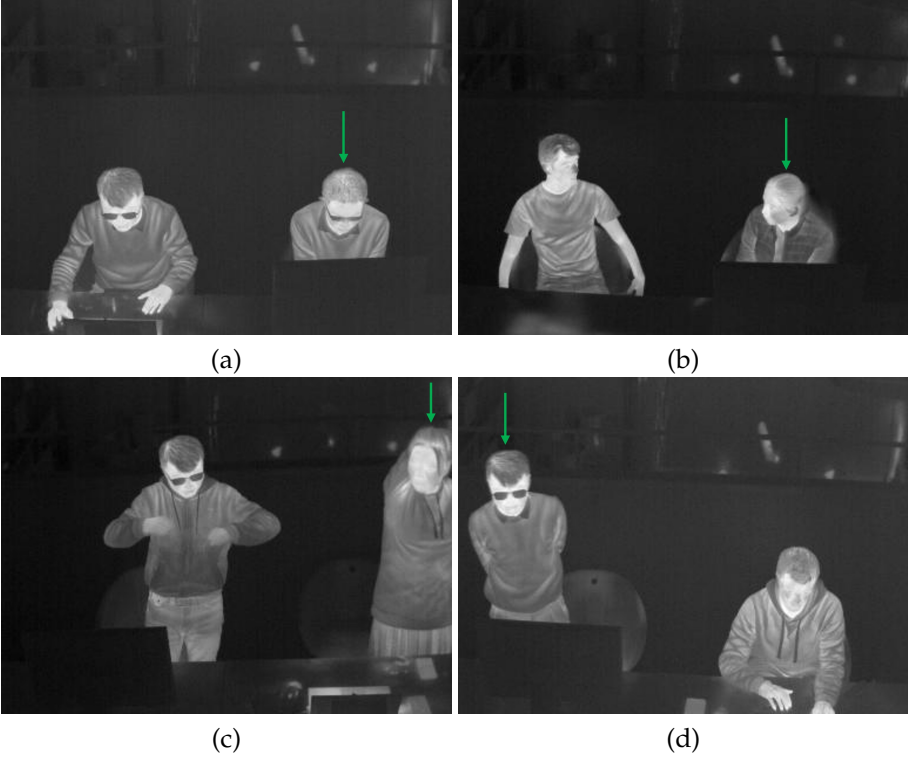


Fig. C.2: Persons difficult to predict clothing type due to occlusion. They are pointed by the green arrows. (a) The right person is partly occluded by the monitor. (b) The right person is partly occluded by the monitor. (c) The right person moves the arms out of the scene. (d) The left person hides the arms behind the torso.

that each person must be tracked across time. To this end, we adopt DeepSort. This tracker receives the image information and YOLOv5-predicted detections, and then decides which tracking ID a detection should be associated to. Like Fig. C.1 shows, DeepSort can use the detected bounding box information in the $(t - 1)$ th frame $(x_{i,t-1}, y_{i,t-1}, w_{i,t-1}, h_{i,t-1})$ indicating the i th box's top-left coordinates, width, height, respectively) to infer the location of the same object in the t th frame in the form of $x'_{i,t}, y'_{i,t}, w'_{i,t}, h'_{i,t}$ by Kalman filter. At the same time, DeepSort extracts and saves the deep features of the object as its appearance information. In this way, two similarity metrics (location and appearance) can be calculated, based on which each detected person can be linked to a specific identity thus making the same person be tracked with a consistent ID over time.

The reason why this DeepSort-by-YOLOv5 paradigm is chosen and applied to such a specific research field is explained further below. The data we

use is in a thermal mode having significantly fewer details compared with its RGB counterpart. This makes the reuse of such limited details/features extremely important. Compared with other detectors, YOLOv5 introduces PANet (Path Aggregation Network) [57] as its neck, making the deeper layers access to the lower-layer features much more efficiently, so the thermal features are well reused. When it comes to the tracking part, the Maximum Age strategy in DeepSort that deletes a track only when it is not associated to any detection more than A_{max} frames can guarantee a consistent ID with the existence of a few false negatives (FN) from YOLOv5. The Tentative Track strategy in DeepSort which confirms a track only after it is associated with detection in three continuous frames also guarantees that occasional false positives (FP) from YOLOv5 have no severe influence on the output. That is to say, this tracking-by-detection framework smooths the direct output from a detector by filtering the undesired consequences of FN and FP, making both the detector and the tracker benefit each other. Additionally, the low complexity and real-time performance of DeepSort fit well the relatively simple scene in our case compared with other cases like pedestrians/vehicles tracking in autonomous driving assistance systems.

Overall, this design not only locates and tracks each individual with a consistent ID in the scene, but also predicts the person's clothing and posture status simultaneously that directly influence I_{cl} and M estimation.

3.2 I_{cl} Estimation

I_{cl} estimation relying on lookup tables in ISO standards [8–10, 12] and updated clothes databases [58, 59] can be a fast solution for laboratory studies, but it is unfeasible to use such a scheme in real applications due to reasons: (i) looking up the I_{cl} value for a person needs extra manual work which is tedious and expensive; (ii) if this look-up task is expected to be done automatically, the solution must have the ability to recognize hundreds of different garment combinations that vary in materials and number of layers as the latest research has revealed the significant importance of them in thermal comfort [2], which is far beyond the capability of existing algorithms.

Therefore, to realize automated estimation, we go another way—using the difference between the skin temperature T_s and the clothes temperature T_c to calculate I_{cl} . This method is intuitive since the difference between T_s and T_c explicitly reveals the heat insulation of clothes to isolate the bare skin from the environmental air. The larger the temperature difference, the higher the clothing insulation rate.

To get T_s and T_c for each individual, the person's skin region R_s and clothing-covered region R_c need to be differentiated from each other. Empirically, R_s includes face, hands, and neck; R_c includes shoulders, torso, and upper arms. However, in daily life, accessories (hat, glasses, scarf, watch,

etc.), spontaneous behaviors (lower one's head, turn one's face away, hide one's arm behind the torso, etc.), and inevitable occlusions by things in front make many body parts be detected unreliably and even totally invisible. After considering such situations, this research counts the lower arms (the middle point of the elbow and wrist) for short-sleeved clothes and the nose area as R_s , and the elbows for long-sleeved clothes and the shoulders as R_c . These regions are also widely used heat receptors in thermal comfort research [35, 54–56]. Fig. C.3 illustrates R_s in green crosses and R_c in red crosses on four images.

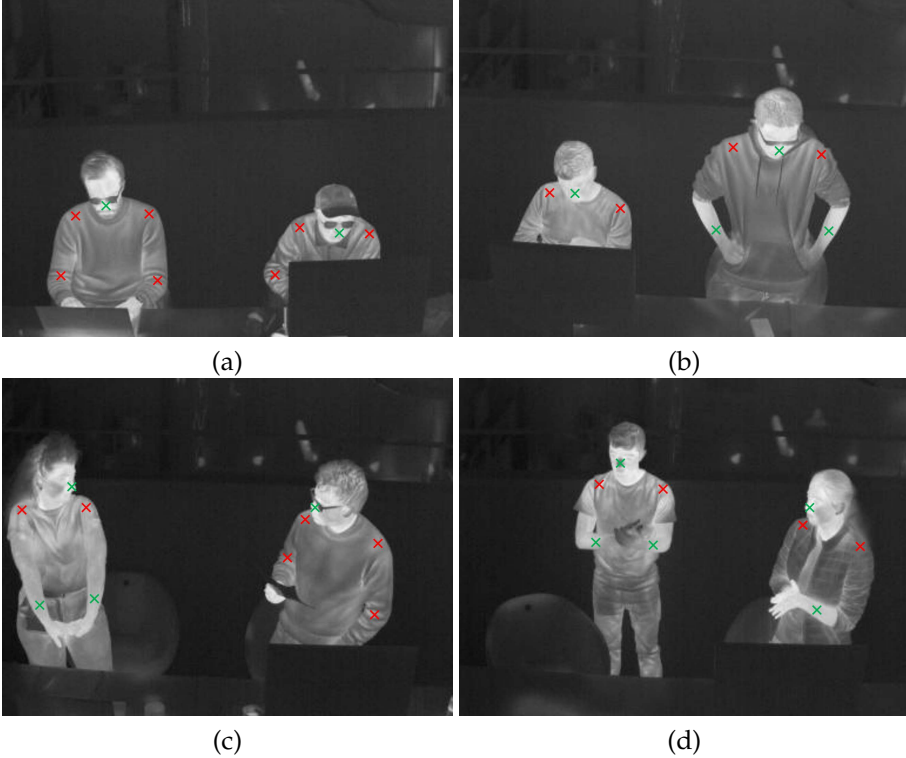


Fig. C.3: Skin region R_s and clothing-covered region R_c . R_s in green crosses and R_c in red crosses. (a)-(d) illustrate persons doing different tasks in different poses.

To locate these body parts, we employ OpenPose [60]—a 2D pose estimation tool. OpenPose has a robust ability against occlusions to detect key body points. The level of the ability against occlusions is determined by a parameter called confidence threshold which means that only the detected key point whose confidence score is higher than the threshold will be counted as the output. The higher threshold, the lower the level of ability against occlusions but the higher accuracy of detection; the lower threshold, the higher-level

3. Methodology

ability against occlusions but more false positives. This can be shown in Fig. C.4 which draws the detected key body points by OpenPose with different confidence thresholds of 0.1, 0.3, 0.5, and 0.7.

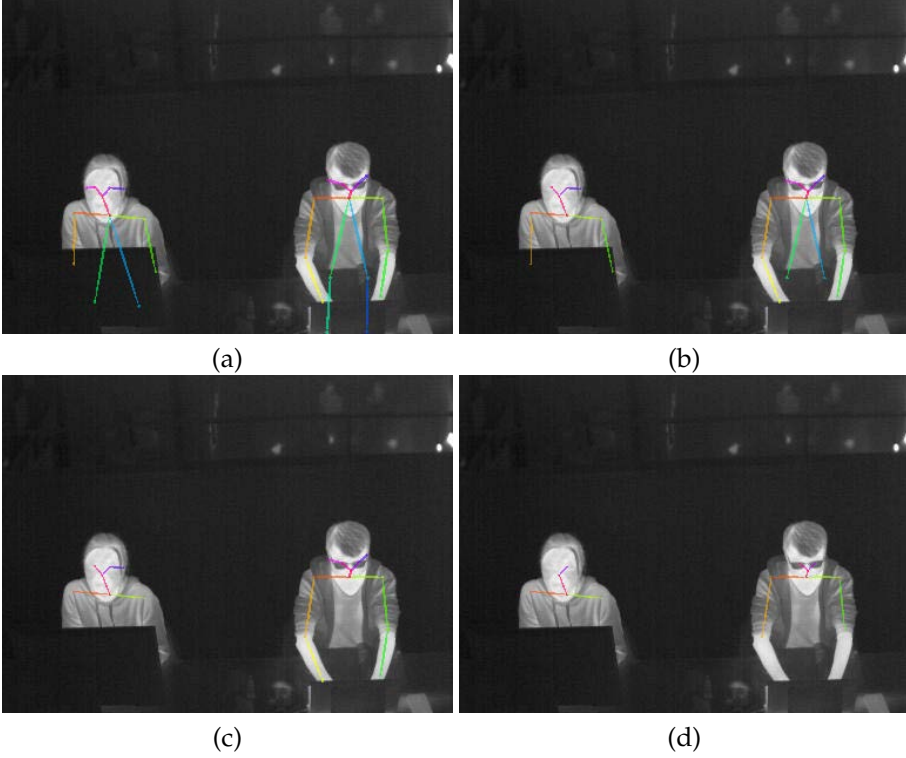


Fig. C.4: Detected key body points by OpenPose with different confidence thresholds. (a) Threshold of 0.1. (b) Threshold of 0.3. (c) Threshold of 0.5. (d) Threshold of 0.7. For better visualization, each key point is the end of the colorful line segment.

Since the detected key points are representations of R_s and R_c and thus directly related to T_s and T_c , a higher accuracy instead of the ability against occlusions is much more important. Like in Fig. C.4(a) and Fig. C.4(b), the detected elbows of the left person are in fact in the computer monitor region; the result in Fig. C.4(c) is more accurate, but the detected wrists of the right person are in the laptop region which will influence the lower arm localization in R_s . These preliminary trials inspire us to set the confidence threshold as high as possible, but a too high threshold produces more missing detections. Therefore, our work uses 0.6 as the threshold in the entire research which has been proved as an effective parameter in the experimental part Section 4.3. To further decrease the influence of miss detections, an accumulation strategy of all the detected key points within a duration like

five minutes is introduced since a person's clothes status is not changed very frequently, which at the same time filters out potential noises.

Another thing worth mentioning is that although OpenPose detects key body points for each person, it has no function of multi-person tracking, and hence our tracking-by-detection framework is still necessary.

In mathematics, based on the recognized sleeves status and OpenPose-predicted key body points, the skin region R_s and the clothing-covered region R_c are determined, both of which are a set of pixel coordinates (x, y) in the image plane like Equation (C.1) and (C.2).

$$R_s = \left\{ (x_t^{1s}, y_t^{1s}), (x_t^{2s}, y_t^{2s}), \dots, (x_{t+1}^{ms}, y_{t+1}^{ms}), \dots, (x_{t+itv-1}^{ns}, y_{t+itv-1}^{ns}) \right\} \quad (C.1)$$

$$R_c = \left\{ (x_t^{1c}, y_t^{1c}), (x_t^{2c}, y_t^{2c}), \dots, (x_{t+1}^{mc}, y_{t+1}^{mc}), \dots, (x_{t+itv-1}^{nc}, y_{t+itv-1}^{nc}) \right\} \quad (C.2)$$

In the equations, the subscript $(t, t+1, t+itv-1)$ refers to the index of each frame within a time period of itv frames; the superscript $(1_s, 2_s, m_s, n_s, 1_c, 2_c, m_c, n_c)$ refers to the index of each detected key point. So in the consecutive itv frames there are n_s and n_c key points detected in R_s and R_c , respectively.

The thermal camera we use is Xenics Gobi-384-GigE that can visualize a thermography of the scene it captures and measure the temperature of each pixel within the image with an accurate resolution of 0.08°C . Therefore, temperatures of the detected key points $(T_{1s}, T_{2s}, \dots, T_{n_s})$ in R_s and $(T_{1c}, T_{2c}, \dots, T_{n_c})$ in R_c are easily read from the camera. Then an average calculation of the temperature values $(T_{1s}, T_{2s}, \dots, T_{n_s})$ and $(T_{1c}, T_{2c}, \dots, T_{n_c})$ gets T_s and T_c , respectively.

As long as T_s and T_c of each individual are calculated, the person's I_{cl} can be estimated by:

$$I_{cl} = \frac{1}{0.155 \cdot h} \cdot \frac{T_s - T_c}{T_c - T_0} \quad (C.3)$$

where h equals to 8.6 referring to human's heat transfer coefficient; T_0 is the operative temperature considering both the air temperature and the mean radiation temperature, so here it is calculated by the average temperature of the background region in each frame. This calculation comes from [35] according to [10, 61], and all the temperatures T_s , T_c , and T_0 are in degrees Celsius. We claim that our emphasis is the OpenPose strategy for localizing R_s and R_c to get T_s and T_c , based on which any I_{cl} calculation method can be applied.

3.3 M Estimation

In this part, we first propose three vision-based features to represent each person's activity intensity, based on which M is estimated.

Three Vision-Based Features

Though M can be estimated by a person's key posture or activity type listed in ISO standards [8, 9, 11, 12] and updated databases [62, 63], this is a rough estimation in many cases, since we have observed that different people tend to have different activity intensities for the same posture. For example, some people will do a bit of stretching when standing up while others may just stand still. Therefore, a more accurate and dynamic M estimation is expected. This is done by computing three vision-based features—a person's bounding box changes in two aspects (location and scale) and the optical flow intensity within the bounding box, over a few seconds like 10 s (210 frames) in our case. Here, the choice of 10 s comes from an observation that it takes similar durations for a smart bracelet to monitor a user's heartbeats and blood oxygen content—two human physiological signals indicating the M value. This three-feature idea is motivated by that: the bounding box location change captures the general body movement; the bounding box scale change captures the motion of limbs; the optical flow intensity within the box captures the subtle movement that the box changes may ignore.

To realize this, for the location change of a certain person's bounding boxes during 10 s (210 frames), the center coordinates (c_x, c_y) of the person's bounding box in each frame is drawn as a point in a 2D plane, and totally the 210 2D points form a cluster-shaped pattern. The more spread out the points are, the larger the general body movement is. The degree of spread can be approximated by fitting an ellipse to the cluster and then calculating the area of this ellipse. In mathematics, first, the covariance matrix of the vector V_{cx} (composed of the horizontal coordinates of the 210 points) and the vector V_{cy} (composed of the vertical coordinates of the 210 points) is computed, and then the two eigenvalues of the covariance matrix are computed, at last, the multiplication of these two eigenvalues represents the area of the ellipse.

For the scale change of a certain person's bounding boxes, after translating the 210 bounding boxes from 210 frames, they will have the same center at the origin, and then the upper-right coordinates (u_x, u_y) of each bounding box represents its scale. Similarly, the 210 upper right points form a cluster in a 2D plane, and the area of the ellipse fitting to the cluster will represent the scale change across time. The larger the area, the larger movement of limbs.

When it comes to the optical flow intensity in a person's bounding box from the t th to $(t + itv - 1)$ th frame (itv equals to 210 here), for each frame

two optical flows in horizontal and vertical directions are extracted by the TV-L1 algorithm [64] realized in a tool called MMAAction [65]. Each optical flow is saved as an 8-bit image in which pixels with a grayscale value of 127 represent no movement while these pixels with grayscale values farther away from 127 represent larger movements. Therefore, within a duration of itv frames, a person's optical flow intensity I_{xy} is calculated by:

$$I_{xy} = \frac{\sum_{\tau=t}^{\tau=t+itv-1} I_{xy}^{\tau}}{itv} \quad (C.4)$$

$$I_{xy}^{\tau} = \sqrt{(I_x^{\tau})^2 + (I_y^{\tau})^2} \quad (C.5)$$

$$I_x^{\tau} = \frac{\sum_{(x,y) \in box_{\tau}} |f_{hrz}^{\tau}(x,y) - 127|}{\sum_{(x,y) \in box_{\tau}} 1} \quad (C.6)$$

$$I_y^{\tau} = \frac{\sum_{(x,y) \in box_{\tau}} |f_{vtc}^{\tau}(x,y) - 127|}{\sum_{(x,y) \in box_{\tau}} 1} \quad (C.7)$$

where τ indicates the frame index; I_{xy}^{τ} is the person's optical flow intensity in the τ th frame; I_x^{τ} and I_y^{τ} are the person's optical flow intensity in the horizontal and vertical directions in the τ th frame, respectively; (x,y) is any pixel in the optical flow; box_{τ} is the bounding box region of the person in the τ th frame; f_{hrz} and f_{vtc} mean the two optical flows in the horizontal and vertical directions, respectively. In Equations (C.6) and (C.7), the number of pixels in the bounding box is acted as the denominator to normalize the influence of the size of the box.

In this way, the three features (bounding box location change, bounding box scale change, optical flow intensity) representing an individual's activity intensity are acquired. A visualization showing the bounding box location change by a cluster of 210 2D points/circles, the bounding box scale change also by a cluster of 210 2D points/circles, and the optical flow intensity within the bounding box in each frame from a duration of 210 frames are in Fig. C.5, in which ID 1 person is standing with very limited movements while ID 2 person is standing and stretching with large movements. This figure intuitively illustrates that the larger body movements of ID 2, the more spread out the points/circles in Fig. C.5(d) and Fig. C.5(f), and the larger optical flow intensity in Fig. C.5(h).

M Estimation from the Three Features

In real life, persons may have various activities which are unrealistic to be analyzed accurately. However, for an office environment, staff usually have scheduled routines and thus relatively fixed behaviors. Generally, the sitting

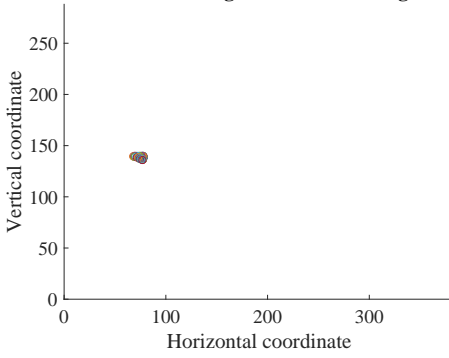
3. Methodology



(a)

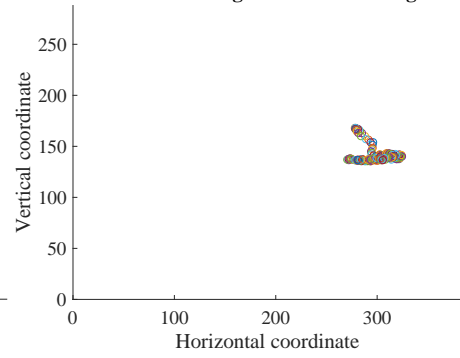
(b)

ID1: Bounding box location change



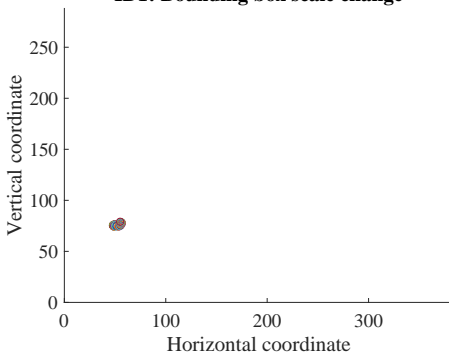
(c)

ID2: Bounding box location change



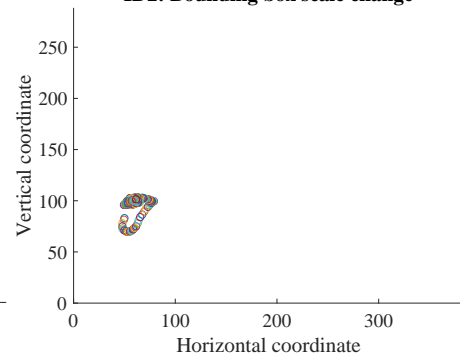
(d)

ID1: Bounding box scale change



(e)

ID2: Bounding box scale change



(f)

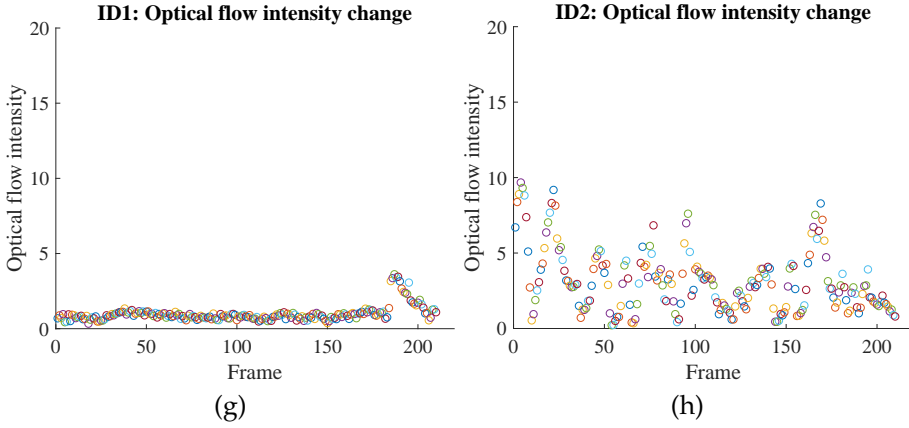


Fig. C.5: Using bounding box changes in location and scale, and the optical flow in the bounding box to represent an individual's activity intensity. (a) and (b): ID 1 person is with small movements and ID 2 person is with large movements. (c) and (d): Bounding box location change of ID 1 person and ID 2 person, respectively. (e) and (f): Bounding box scale change of ID 1 person and ID 2 person, respectively. (g) and (h): Optical flow intensity change of ID 1 person and ID 2 person, respectively.

staff are typing the keyboard, reading, taking notes, sorting through files, chatting with colleagues, online meetings, etc. And the standing staff are also occupied by the same tasks but may be involved with some walking or body stretching. This prior knowledge is such important that it gives a metabolic rate range from which each individual's M varies.

Therefore, with the above prior knowledge of standard office behaviors, by referring Table A.1 and Table A.2 in ISO 8996 [11], the CBE (Center for the Built Environment) thermal comfort tool [66], and the 2011 compendium of physical activities tables [63, 67], the usual metabolic rate range of a sitting office staff is quite narrow from 58 W/m^2 (1.0 MET) to 87 W/m^2 (1.5 MET), while a standing staff's metabolic rate usually varies from 75 W/m^2 (1.3 MET) to 174 W/m^2 (3.0 MET). According to the CBE thermal comfort tool, the slight M change of a sitting person within the range $[58 \text{ W/m}^2, 87 \text{ W/m}^2]$ has a mild influence on his or her thermal sensation, while the M change within the much larger range of a standing person significantly influences the thermal feeling. This result inspires us to use a middle value of 72.5 W/m^2 to represent a sitting office staff's M for simplicity and generalization which also relieves the three-feature extraction for him or her, but we need to specifically define a standing person's M from his or her dynamic activity intensity situation represented by the three vision-based features.

To map such features to a value of M , a classification idea is introduced. Similar to Table A.2 in ISO 8896 where metabolic rates from 55 W/m^2 to more than 260 W/m^2 are categorized into resting, low, moderate, high, and very

4. Experiments

high levels, we decide to categorize the metabolic rate of a standing office staff into low, moderate, and high levels. Specifically, a low level means standing with very limited movements or transient spontaneous movements (standing quietly in a line, reading, using a cellphone, normally chatting, etc.); a moderate level means standing with spontaneous but lasting movements (natural and small paces, limbs movements, head movements, discussing with gestures, etc.); a high level means standing with significant movements usually indicating intentional actions like sustained location changes by walking, constant trunk movements to stretch/relax the body, etc.

It is extremely important that the three levels do not mean there are only three options for the M value. Instead, for a person's activity intensity, there are three classification probabilities P_l , P_m , and P_h indicating the possibilities of being viewed as low, moderate, and high level, respectively. Based on P_l , P_m , and P_h , the person's final M is estimated by:

$$M = P_l \cdot M_l + P_m \cdot M_m + P_h \cdot M_h \quad (\text{C.8})$$

where M_l , M_m , and M_h are the lower boundary, the middle value, and the upper boundary of a standing person's M , that are, 75 W/m², 125 W/m², and 174 W/m², respectively.

To realize this solution, the classification probabilities P_l , P_m , and P_h are in need. With only three features describing a person's activity intensity within a few seconds as the input, a simple and flexible classification model instead of a CNN can be used. So, in this study, several lightweight models are employed and the random forest model works best. The training and testing details are in Section 4.4.

In summary, the proposed M estimation method has several advantages: (i) the three explicitly-extracted features can guide the metabolic rate estimation efficiently, considering that the features automatically extracted by a learning method are relatively difficult to anticipate and thus may potentially fail for a specific task; (ii) the three features are really low dimensional, making it possible to use lightweight machine learning classifiers which are flexible to be integrated into the whole system; (iii) the probability-weighted summation (Equation (C.8)) makes the estimated M continuously change in a range, which not only fits the real-life scenario than limited and discrete choices in existing methods but also avoids the very difficult annotation if a regression model is adopted.

4 Experiments

In this part, we first introduce the information of the dataset we collected from a multi-person environment, and then the proposed tracking-

by-detection module, I_{cl} estimation module, and M estimation module are evaluated.

4.1 Dataset Information

There is no available public dataset for visual analysis of I_{cl} and M in a multi-person environment. We, therefore, collected such a dataset in December 2020 in Denmark. During the collection, two persons were sitting or standing with different types of clothes in a typical office environment where the indoor temperature and humidity were 22 °C and 32%, and they were encouraged to behave naturally. That means, typing the keyboard, texting with cellphones, chatting with each other, reading, stretching the body to relax, and others were captured in the collected videos. The horizontal distance between the camera and persons is around 3.5 meters, and the vertical distance between the camera and the ground is around 2.7 meters. In this way, ten subjects contributed to 114 videos with each video’s length about 2000 frames by using a thermal camera (Xenics Gobi-384-GigE whose sensor size is 384×288).

4.2 Evaluation of the Tracking-by-Detection Module

The tracking-by-detection (DeepSort-by-YOLOv5) module needs a well-trained human detector to detect persons in six categories mentioned before in Table C.3. To train YOLOv5, from the dataset we sampled one frame every 50 frames for annotation and thus 5263 frames are selected in which each person’s bounding box and category are labeled. These 5263 images are then divided into a training set (4467), validation set (362), and testing set (434) to guarantee that subjects in the testing set never exist in the training set and validation set for a fair evaluation. Additionally, we selected and labeled 832 images from a single-person thermal dataset from [34] to increase the amount and diversity of the training set. The detailed information of the data to train and evaluate YOLOv5 is listed in Table C.4. Accordingly, the 15 videos from which the 434 testing images are sampled are used to evaluate the whole DeepSort-by-YOLOv5 framework.

With a desktop equipped with Windows 10, CUDA (Compute Unified Device Architecture) 10.2, PyTorch 1.7.1, and one NVIDIA 2080Ti GPU (Graphics Processing Unit) card, the YOLOv5m version [41] is finetuned with the learning rate 0.0075 and stops at the 200th epoch at which the training loss is not decreasing any more. Other settings remain the same with the released YOLOv5m. The best model on the validation set is performed on the testing set and then achieves a mAP_{50} (Mean Average Precision) of 89.1% over six categories. Specifically, the AP_{50} rates of LongSit, ShortSit, OclSit, LongStand, ShortStand, and OclStand are 98.8%, 90.0%, 95.5%, 98.5%, 99.5%, and 52.5%,

4. Experiments

Table C.4: Detailed information of the data to train and evaluate YOLOv5.

-		Training	Validation	Testing
Number of images		5299	362	434
Number of persons	LongSit	2099	172	22
	ShortSit	1615	29	157
	OclSit	828	274	92
	LongStand	2280	140	149
	ShortStand	2735	100	443
	OclStand	254	9	2

respectively. The AP_{50} drop in OclStand is due to the data imbalance problem. There are less than 300 images having OclStand persons in the training set, and there are only two images having Oclstand persons in the testing set (see Fig. C.6). In Fig. C.6, persons with bounding box SSD (ShortStand), SS(ShortSit), and OSD (OclStand) are categorized correctly, while the one with box LSD (LongStand) is categorized wrongly since the person’s sleeve status is unknown and thus should have been recognized as OclStand (OSD).

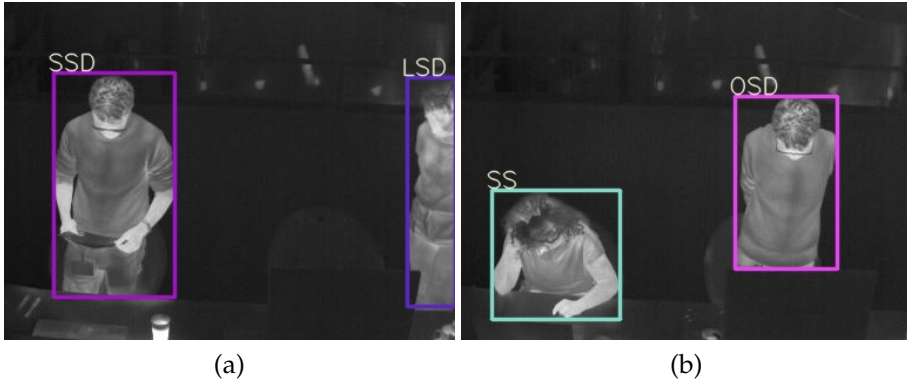


Fig. C.6: Detection results on two test images with OclStand (OSD) persons in them. (a) The right person is wrongly categorized as LongStand (LSD). (b) Both persons are detected and categorized correctly.

With the same hardware and software platforms, DeepSort-by-YOLOv5 runs on the 15 testing videos without further fine-tuning of the tracker itself. There are a total of 44,077 ground truth persons, 206 false negatives, 16 false positives, and 0 ID-switch in the 15 videos, which achieves an average MOTA (Multiple Object Tracking Accuracy) of 99.5% and the lowest MOTA of an individual video is 93.7%. Fig. C.7 shows four sampled tracking results. The eight persons from left to right in Fig. C.7 are in category ShortSit, LongStand, ShortSit, LongStand, ShortSit, OclSit, ShortStand, and

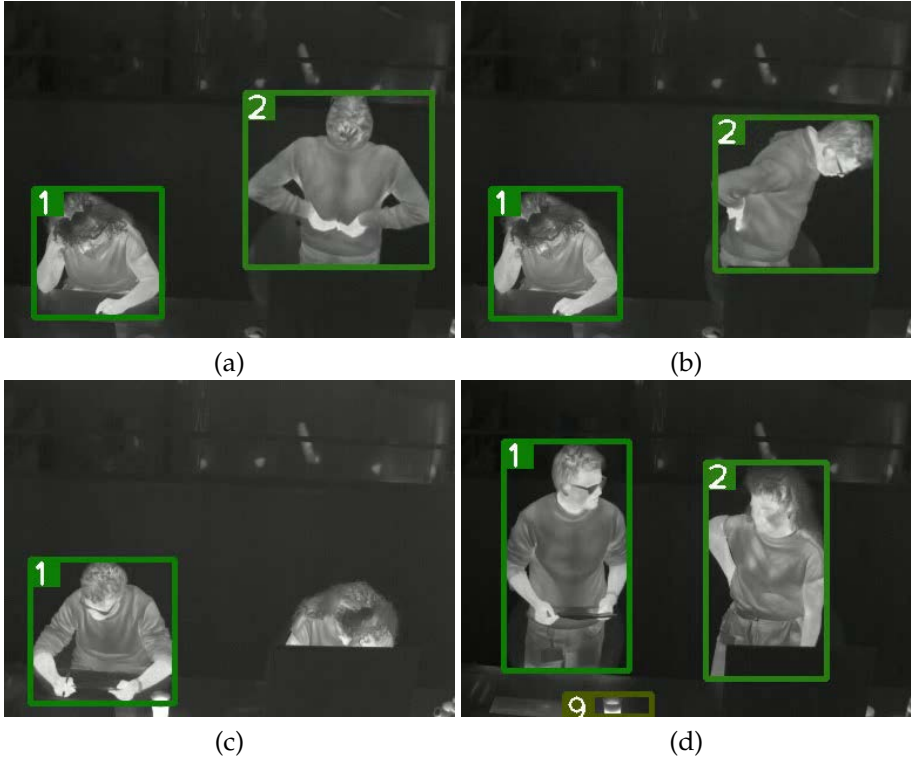


Fig. C.7: Sampled tracking results on the testing set. No false positive and false negative in (a) and (b). One false negative in (c). One false positive in (d). The numbers indicate tracked ID numbers.

ShortStand, respectively. Fig. C.7(a) and Fig. C.7(b) are near frames from a video, and both persons are well tracked though the person with ID 2 is moving intensely. The false negative in Fig. C.7(c) is because there is no similar situation in the training set that a person is occluded so severely. The mug with hot coffee in Fig. C.7(d) has a similar temperature distribution as humans, which leads to the false positive.

In summary, the proposed DeepSort-by-YOLOv5 module achieves a mAP_{50} rate of 89.1% and a MOTA rate of 99.5% on the testing data. As this is the first work on multi-person analysis in terms of clothing and activity status recognition for thermal comfort, a direct comparison with other works is not possible. Instead, we refer to the latest performance of human detection/tracking on other thermal databases as an indirect comparison. Work [51] shows that the mAP_{50} values are from 62.0% to 96.0% on benchmark databases with different difficulties like OSU, KAIST, VOT-TIR2015, etc. Work [68] shows that the MOTA values are from 54.3% to 64.9% with

different trackers on SCUT-FIR pedestrian dataset. These reference results indicate that our results are good enough and thus the proposed method can be included in a real application.

4.3 Evaluation of the I_{cl} Estimation Module

The I_{cl} estimation closely depends on the skin temperature T_s and clothes temperature T_c acquisition, which is bridged by the localization of skin region R_s and clothing-covered region R_c via OpenPose. Therefore, this evaluation first looks at the efficacy of applying OpenPose to our dataset.

4901 images are used to examine OpenPose’s performance. These 4901 images come from the 5263 annotated images for YOLOv5 but do not include the images where persons are wearing masks due to coronavirus restrictions. Such an evaluation set is evenly sampled from the 114 collected videos, guaranteeing comprehensiveness and fairness. The evaluation protocols are: (i) the OpenPose tool is not finetuned with our thermal dataset, and the confidence threshold is set as 0.6 as mentioned in Section 3.2; (ii) only these key points that influence R_s and R_c localization are checked, i.e., nose, shoulders, elbows, and wrists; (iii) any frame with even only one wrongly detected key body point is counted as one error frame, to make the evaluation strict and conservative.

After a frame-by-frame check, there are 187 error frames out of the whole 4901 frames, indicating an accuracy of 96.2%. We found that there are two types of representative errors—nose detected in the hair region due to a lowering head (Fig. C.8(a)) and nose detected in the background region due to a turned side face (Fig. C.8(b)). The good point is that with the average computation within a few minutes to get T_s and T_c , the influence of these errors can be eliminated effectively, and of course, a higher confidence threshold can further reduce such errors if needed.

Therefore, the efficacy of applying OpenPose to our multi-person thermal scenario to locate R_s and R_c is verified. The performance surpasses that of applying OpenPose to a controlled single-person thermal environment [35] and applying OpenPose to RGB MPII dataset [69], further proving the feasibility of our strategy relying on OpenPose.

Based on the above acquired R_s and R_c , here we calculate the T_s and T_c , and then estimate the I_{cl} value. Since an individual I_{cl} estimation also involves the human tracking part, we use the testing videos for the tracking module to evaluate this I_{cl} estimation module too. From the testing videos, a female wearing a lightweight T-shirt is acting as the subject to be researched, because there is an available reference for her clothes type in the ISO tables so that we can make a comparison. And thus, two videos including various situations where the female is sitting, standing, reading, writing, typing the keyboard, chatting, and drinking coffee (some frames are shown in Fig. C.9)

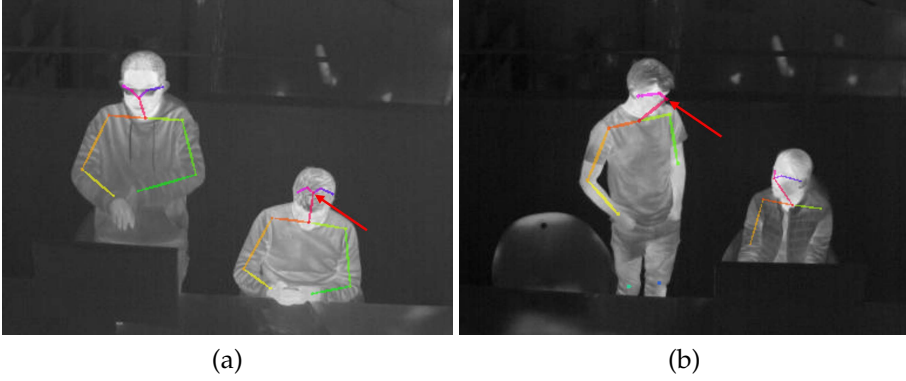


Fig. C.8: Two representative error frames with red arrows pointing to the wrongly detected noses. (a) Nose is detected in the hair region. (b) Nose is detected in the background region.

go through our methodology pipeline to get her I_{cl} . In one video consisting of 1477 frames (70 seconds), 3326 skin points and 2849 clothes points are detected for the female, from which the T_s and T_c are calculated as 34.67 °C and 33.32 °C, respectively. Together with the T_o as 24.96 °C, the female's I_{cl} is estimated as 0.1220 clo. In the other video of 1536 frames (73 seconds), 2496 skin points and 2502 clothes points are detected for the female; the resultant T_s is 34.73 °C and T_c is 33.48 °C; together with the T_o as 25.58 °C, the female's I_{cl} is estimated as 0.1182 clo.

From above calculation, we find that: (i) within a time period like more than 1 minute, the accumulated detected points in R_s and R_c are way enough for an accurate T_s and T_c calculation as the potential noises can be filtered out efficiently; (ii) the estimated I_{cl} values of 0.1220 clo and 0.1182 clo are quite similar, revealing the stability and robustness of the method; (iii) the reference value of the female's I_{cl} is 0.09 clo to 0.15 clo from Table B.1 in ISO 9920 [10], showing the consistency of our method with the international standards, and proving the feasibility of the proposed method.

4.4 Evaluation of the M Estimation Module

This subsection evaluates the effectiveness of the M estimation based on the three extracted vision features, specifically for a standing person. As this estimation is a probability-weighted summation, measuring the accuracy of the classifier is the key.

Therefore, by dividing the 114 collected videos into small clips of 10 seconds and then extracting the three vision features for each standing person in these clips, 315 sets of the three features are used as the training data to help the classifier learn the ability to category each person's activity intensity into low, moderate, or high level, and another 68 sets are used as the testing

4. Experiments

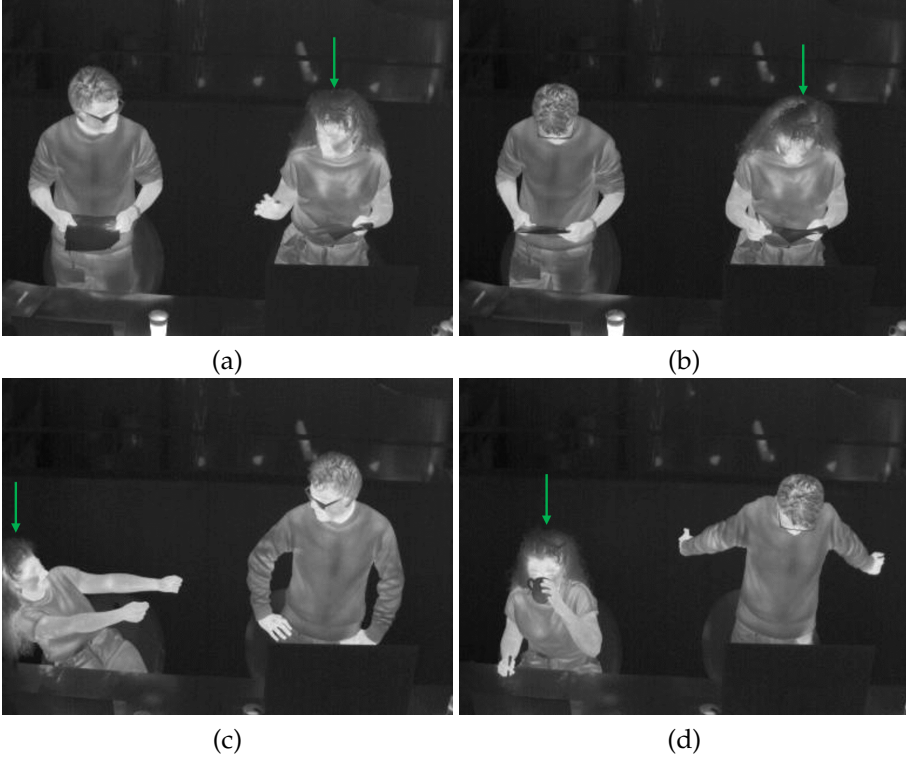


Fig. C.9: The female to be researched is pointed by the green arrow. (a) Chatting. (b) Reading. (c) Chatting with gestures. (d) Drinking.

data to evaluate the classifier’s performance.

During the phase of preparing the training and testing data—annotating a standing person’s activity intensity level, we met another dilemma that frequently happens in the real world—there are always the situations where a person’s movement is mixed with transient, lasting, mild, or intensive movements within a short period which makes it very difficult to label the intensity level. Therefore, these difficult cases are not included in the training/testing sets to not confuse the classifier. From the positive side, this situation further indicates the strength of our probability-weighted summation strategy that makes the estimated M a continuous value.

To avoid being one-sided, three widely-used classifiers—KNN (K-NearestNeighbor), SVM (Support Vector Machine), and RF (Random Forest) are used. The parameters and performances of the three classifiers are listed in Table C.5, in which each parameter is tuned by grid searching using the training data and the meaning of each parameter is explained in the scikit-learn library [70]. These accuracy values in Table C.5 prove that the three

features are good representations of a person’s activity intensity, and thus the M estimation from them by a classifier’s probability-weighted summation is also reasonable. And then we decide to use RF as the classifier for M estimation due to its best performance on the testing data.

Table C.5: The parameters and performances of the used three classifiers.

Classifier	Parameters	Training accuracy	Testing accuracy
KNN	metric='manhattan', weights='distance', n_neighbors=13	100%	92.7%
SVM	C=50, kernel='rbf', gamma='scale'	83.5%	88.2%
RF	max_depth=2, random_state=0	95.6%	95.6%

Based on RF’s classification probabilities P_l , P_m , and P_h , by Equation (C.8), the M values of a same standing person with two totally different activity intensities are estimated. The person is shown in Fig. C.10, in which Fig. C.10(a) is a frame from a clip where the standing person is normally chatting with many gestures, and Fig. C.10(b) is a frame from another clip where the standing person is stretching his body like doing Pilates. For them, our method outputs the estimated M values of 99 W/m^2 and 170 W/m^2 , respectively, which are very similar to the reference values of 104 W/m^2 (CODE 09050 in [67]) and 174 W/m^2 (CODE 02105 in [67]), further proving the feasibility and usability of the proposed M estimation module.

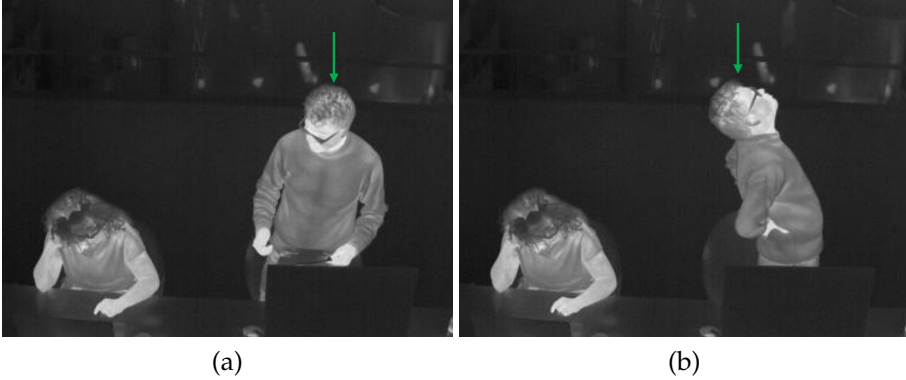


Fig. C.10: The standing person to be researched for M estimation is pointed by the green arrow. (a) Normally chatting but with many gestures. (b) Stretching body like doing Pilates.

4.5 Application in Thermal Comfort Assessment

From all the above evaluations, the proposed method indeed has the ability to estimate individual I_{cl} and M across time for each person in a room. With these two dynamic personal factors and the other four environmental factors easily measured from sensors, a thermal comfort model like Fanger's model [6, 7] can calculate individual thermal comfort sensation to see if the person feels hot, cold, or satisfied with the indoor environment. Although occupants may have different thermal feelings at the same time, by regulating the indoor microclimate in separate local regions, it is possible to achieve varied thermal conditions that respond to the different subjective thermal states. Moreover, the used thermal camera instead of an RGB camera, the computation in a local device, and the erasing function of captured image information as long as I_{cl} and M are estimated will make the whole processing pipeline privacy-friendly.

5 Conclusions and Future Work

This paper proposes a contactless method to estimate each person's clothing insulation rate I_{cl} and metabolic rate M dynamically by use of a thermal camera, in an uncontrolled multi-person indoor environment.

Specifically, the method composes of a tracking-by-detection (DeepSort-by-YOLOv5) module to track each person and recognize his or her clothing status and key posture simultaneous, a key body points detection module to measure the skin temperature and clothes temperature for I_{cl} estimation, and a random forest classifier module to categorize each individual's activity intensity into different levels for M estimation. All three modules are evaluated with a new multi-person thermal dataset, verifying that the methodology is robust to be applied in real-life applications for individual thermal comfort assessment.

The future work will be to include this research into such an application to facilitate thermal comfort control systems for lower energy waste and higher working comfort in an office building.

References

- [1] EIA, "2012 commercial buildings energy consumption survey data," <https://www.eia.gov/consumption/commercial/data/2012/>, last accessed: August, 2021.
- [2] Desalegn Atalie, Pavla Tesinova, Melkie Getnet Tadesse, Eyasu Ferede, Ionuț Dulgheriu, and Emil Loghin, "Thermo-physiological comfort

- properties of sportswear with different combination of inner and outer layers,” *Materials*, vol. 14, no. 22, pp. 6863, 2021.
- [3] Hafiz Muhammad Kaleem Ullah, Joseph Lejeune, Aurélie Cayla, Mélanie Monceaux, Christine Campagne, and Éric Devaux, “A review of noteworthy/major innovations in wearable clothing for thermal and moisture management from material to fabric structure,” *Textile Research Journal*, p. 00405175211027799, 2021.
 - [4] Yangyang Peng, Fengxin Sun, Caiqin Xiao, Mohammad Irfan Iqbal, Zhenguo Sun, Mingrui Guo, Weidong Gao, and Xiaorui Hu, “Hierarchically structured and scalable artificial muscles for smart textiles,” *ACS Applied Materials & Interfaces*, vol. 13, no. 45, pp. 54386–54395, 2021.
 - [5] FL Zhu and QQ Feng, “Recent advances in textile materials for personal radiative thermal management in indoor and outdoor environments,” *International Journal of Thermal Sciences*, vol. 165, pp. 106899, 2021.
 - [6] Poul O Fanger et al., “Thermal comfort. analysis and applications in environmental engineering,” *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.
 - [7] Povl Ove Fanger, “Assessment of man’s thermal comfort in practice,” *Occupational and Environmental Medicine*, vol. 30, no. 4, pp. 313–324, 1973.
 - [8] International Standard Organization, “Ergonomics of the thermal environment — analytical determination and interpretation of thermal comfort using calculation of the pmv and ppd indices and local thermal comfort criteria,” <https://www.iso.org/standard/39155.html/>, last accessed: September, 2021.
 - [9] International Standard Organization, “Ergonomics of the thermal environment — analytical determination and interpretation of heat stress using calculation of the predicted heat strain,” <https://www.iso.org/standard/37600.html/>, last accessed: September, 2021.
 - [10] International Standard Organization, “Ergonomics of the thermal environment — estimation of thermal insulation and water vapour resistance of a clothing ensemble,” <https://www.iso.org/standard/39257.html/>, last accessed: September, 2021.
 - [11] International Standard Organization, “Ergonomics of the thermal environment — determination of metabolic rate,” <https://www.iso.org/standard/34251.html/>, last accessed: September, 2021.
 - [12] ASHRAE, “Thermal environmental conditions for human occupancy,” <http://arco-hvac.ir/wp-content/uploads/2015/11/>

References

- ASHRAE_Thermal_Comfort_Standard.pdf, last accessed: September, 2021.
- [13] Michele De Carli, Bjarne W Olesen, Angelo Zarrella, and Roberto Zecchin, "People's clothing behaviour according to external weather and indoor environment," *Building and Environment*, vol. 42, no. 12, pp. 3965–3973, 2007.
- [14] Jack Ngarambe, Geun Young Yun, and Gon Kim, "Prediction of indoor clothing insulation levels: A deep learning approach," *Energy and Buildings*, vol. 202, pp. 109402, 2019.
- [15] Paulo Matos de Carvalho, Manuel Gameiro da Silva, and João Esteves Ramos, "Influence of weather and indoor climate on clothing of occupants in naturally ventilated school buildings," *Building and environment*, vol. 59, pp. 38–46, 2013.
- [16] Weiwei Liu, Diyu Yang, Xiong Shen, and Peizhi Yang, "Indoor clothing insulation and thermal history: A clothing model based on logistic function and running mean outdoor temperature," *Building and Environment*, vol. 135, pp. 142–152, 2018.
- [17] Hiroki Matsumoto, Yoshio Iwai, and Hiroshi Ishiguro, "Estimation of thermal comfort by measuring clo value without contact.," in *MVA*. Citeseer, 2011, pp. 491–494.
- [18] Maria Konarska, Krzysztof Soltynski, Iwona Sudol-Szopinska, and Anna Chojnacka, "Comparative evaluation of clothing thermal insulation measured on a thermal manikin and on volunteers," *Fibres and Textiles in Eastern Europe*, vol. 15, no. 2, pp. 73, 2007.
- [19] Siliang Lu and Erica Cochran Hameen, "Integrated ir vision sensor for online clothing insulation measurement," 2018.
- [20] Jeong-Hoon Lee, Young-Keun Kim, Kyung-Soo Kim, and Soohyun Kim, "Estimating clothing thermal insulation using an infrared camera," *Sensors*, vol. 16, no. 3, pp. 341, 2016.
- [21] Kyungsoo Lee, Haneul Choi, Hyungkeun Kim, Daeung Danny Kim, and Taeyeon Kim, "Assessment of a real-time prediction method for high clothing thermal insulation using a thermoregulation model and an infrared camera," *Atmosphere*, vol. 11, no. 1, pp. 106, 2020.
- [22] Haneul Choi, HooSeung Na, Taehung Kim, and Taeyeon Kim, "Vision-based estimation of clothing insulation for building control: A case study of residential buildings," *Building and Environment*, p. 108036, 2021.

- [23] Maohui Luo, Xiang Zhou, Yingxin Zhu, and Jan Sundell, "Revisiting an overlooked parameter in thermal comfort studies, the metabolic rate," *Energy and Buildings*, vol. 118, pp. 152–159, 2016.
- [24] Yongchao Zhai, Minghui Li, Siru Gao, Liu Yang, Hui Zhang, Edward Arens, and Yunfei Gao, "Indirect calorimetry on the metabolic rate of sitting, standing and walking office activities," *Building and Environment*, vol. 145, pp. 77–84, 2018.
- [25] Wenjie Ji, Maohui Luo, Bin Cao, Yingxin Zhu, Yang Geng, and Borong Lin, "A new method to study human metabolic rate changes and thermal comfort in physical exercise by co2 measurement in an airtight chamber," *Energy and Buildings*, vol. 177, pp. 402–412, 2018.
- [26] Andrea Calvaresi, Marco Arnesano, Filippo Pietroni, and Gian Marco Revel, "Measuring metabolic rate to improve comfort management in buildings.," *Environmental Engineering & Management Journal (EEMJ)*, vol. 17, no. 10, 2018.
- [27] Mohammad H Hasan, Fadi Alsaleem, and Mostafa Rafaie, "Sensitivity study for the pmv thermal comfort model and the use of wearable devices biometric data for metabolic rate estimation," *Building and Environment*, vol. 110, pp. 173–183, 2016.
- [28] Yuchun Zhang, Xiaoqing Zhou, Zhimin Zheng, Majeed Olaide Oladokun, and Zhaosong Fang, "Experimental investigation into the effects of different metabolic rates of body movement on thermal comfort," *Building and Environment*, vol. 168, pp. 106489, 2020.
- [29] Jeehee Lee and Youngjib Ham, "Physiological sensing-driven personal thermal comfort modelling in consideration of human activity variations," *Building Research & Information*, vol. 49, no. 5, pp. 512–524, 2021.
- [30] Syed Ihtsham-ul-Haq Gilani, Muhammad Hammad Khan, and Muzaffar Ali, "Revisiting fanger's thermal comfort model using mean blood pressure as a bio-marker: An experimental investigation," *Applied thermal engineering*, vol. 109, pp. 35–43, 2016.
- [31] Martin Møller Jensen, Mathias Krogh Poulsen, Thiemo Alldieck, Ryan Godsk Larsen, Rikke Gade, Thomas B Moeslund, and Jesper Franch, "Estimation of energy expenditure during treadmill exercise via thermal imaging," *Medicine and science in sports and exercise*, vol. 48, no. 12, pp. 2571–2579, 2016.
- [32] Rikke Gade, Ryan Godsk Larsen, and Thomas B Moeslund, "Measuring energy expenditure in sports by thermal video analysis," in *Proceedings*

- of the *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 131–138.
- [33] Hooseung Na, Haneul Choi, and Taeyeon Kim, “Metabolic rate estimation method using image deep learning,” in *Building Simulation*. Springer, 2020, vol. 13, pp. 1077–1093.
 - [34] Jinsong Liu, Isak Worre Foged, and Thomas B Moeslund, “Vision-based individual factors acquisition for thermal comfort assessment in a built environment,” in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 662–666.
 - [35] Jinsong Liu, Isak Worre Foged, and Thomas B Moeslund, “Automatic estimation of clothing insulation rate and metabolic rate for dynamic thermal comfort assessment,” *Pattern Analysis and Applications*, pp. 1–16, 2021.
 - [36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.
 - [37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
 - [38] Joseph Redmon and Ali Farhadi, “Yolo9000: better, faster, stronger,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.
 - [39] Joseph Redmon and Ali Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
 - [40] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, “Yolov4: Optimal speed and accuracy of object detection,” *arXiv preprint arXiv:2004.10934*, 2020.
 - [41] Ultralytics, “Yolov5,” <https://github.com/ultralytics/yolov5/>, 2020, last accessed: March, 2021.
 - [42] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

- [43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [44] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [45] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.
- [46] Nicolai Wojke, Alex Bewley, and Dietrich Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.
- [47] Seung-Hwan Bae and Kuk-Jin Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 595–610, 2017.
- [48] Dawei Zhao, Hao Fu, Liang Xiao, Tao Wu, and Bin Dai, "Multi-object tracking with correlation filter for autonomous vehicle," *Sensors*, vol. 18, no. 7, pp. 2004, 2018.
- [49] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy, "Robust multi-modality multi-object tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2365–2374.
- [50] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang, "Retina-track: Online single stage joint detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14668–14678.
- [51] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar, "Thermal object detection in difficult weather conditions using yolo," *IEEE Access*, vol. 8, pp. 125459–125476, 2020.
- [52] Noor Ul Huda, Bolette D Hansen, Rikke Gade, and Thomas B Moeslund, "The effect of a diverse dataset for transfer learning in thermal person detection," *Sensors*, vol. 20, no. 7, pp. 1982, 2020.
- [53] Jinsong Liu, Mark P Philipsen, and Thomas B Moeslund, "Supervised versus self-supervised assistant for surveillance of harbor fronts," in *VISIGRAPP (5: VISAPP)*, 2021, pp. 610–617.

- [54] Ye Yao, Zhiwei Lian, Weiwei Liu, and Chunxiao Jiang, "Measurement methods of mean skin temperatures for the pmv model," *HVAC&R Research*, vol. 14, no. 2, pp. 161–174, 2008.
- [55] Changzhi Dai, Hui Zhang, Edward Arens, and Zhiwei Lian, "Machine learning approaches to predict thermal demands using skin temperatures: Steady-state conditions," *Building and Environment*, vol. 114, pp. 1–10, 2017.
- [56] Andrei Claudiu Cosma and Rahul Simha, "Thermal comfort modeling in transient conditions using real-time local body temperature extraction with a thermographic camera," *Building and Environment*, vol. 143, pp. 36–47, 2018.
- [57] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.
- [58] Jakob Petersson and Amitava Halder, "Updated database of clothing thermal insulation and vapor permeability values of western ensembles for use in ashrae standard 55, iso 7730, and iso 9920," *ASHRAE Transactions*, vol. 127, pp. 773–799, 2021.
- [59] Yin Tang, Zixiong Su, Hang Yu, Kege Zhang, Chaoen Li, and Hai Ye, "A database of clothing overall and local insulation and prediction models for estimating ensembles' insulation," *Building and Environment*, vol. 207, pp. 108418, 2022.
- [60] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.
- [61] Jun Miura, Mitsuhiro Demura, Kaichiro Nishi, and Shuji Oishi, "Thermal comfort measurement using thermal-depth images for robotic monitoring," *Pattern Recognition Letters*, vol. 137, pp. 108–113, 2020.
- [62] Barbara E Ainsworth, William L Haskell, Melicia C Whitt, Melinda L Irwin, Ann M Swartz, Scott J Strath, WILLIAM L O'Brien, David R Bassett, Kathryn H Schmitz, Patricia O Emplainscourt, et al., "Compendium of physical activities: an update of activity codes and met intensities," *Medicine and science in sports and exercise*, vol. 32, no. 9; SUPP/1, pp. S498–S504, 2000.
- [63] Barbara E Ainsworth, William L Haskell, Stephen D Herrmann, Nathanael Meckes, David R Bassett, Catrine Tudor-Locke, Jennifer L

- Greer, Jesse Vezina, Melicia C Whitt-Glover, and Arthur S Leon, "2011 compendium of physical activities: a second update of codes and met values," *Med Sci Sports Exerc*, vol. 43, no. 8, pp. 1575–1581, 2011.
- [64] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, pp. 214–223.
- [65] Dahua Lin Yue Zhao, Yuanjun Xiong, "Mmaction," <https://github.com/open-mmlab/mmaction>, 2019, last accessed: March, 2021.
- [66] Federico Tartarini, Stefano Schiavon, Toby Cheung, and Tyler Hoyt, "Cbe thermal comfort tool: Online tool for thermal comfort calculations and visualizations," *SoftwareX*, vol. 12, pp. 100563, 2020.
- [67] Barbara E Ainsworth, William L Haskell, Stephen D Herrmann, Nathanael Meckes, David R Bassett, Catrine Tudor-Locke, Jennifer L Greer, Jesse Vezina, Melicia C Whitt-Glover, and Arthur S Leon, "2011 compendium of physical activities tables," <https://sites.google.com/site/compendiumofphysicalactivities/compendia>, 2019, last accessed: September, 2021.
- [68] Huanjie Chen, Wenjie Cai, Feng Wu, and Qiong Liu, "Vehicle-mounted far-infrared pedestrian detection using multi-object tracking," *Infrared Physics & Technology*, vol. 115, pp. 103697, 2021.
- [69] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.
- [70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Paper D

Knowing Where to Look: Gender Classification in Thermal Imagery

Jinsong Liu and Thomas B. Moeslund

This is a technical report of ongoing work, 2022

© 2022 Authors

The layout has been revised.

Abstract

For a built environment that thermally satisfies the humans in it, the indoor thermal conditions should change adaptively according to the occupants' thermal sensations and needs. Recently more studies have found that the gender difference has an impact on one's thermal sensation, which has stimulated interest in gender classification. Accordingly, this work investigates this topic in thermal imagery for its privacy-friendly property instead of in more common RGB visible imagery. Considering the lack of details in a thermal image compared with its visible counterpart, data-driven-based CNN classifiers for thermal inputs usually encounter more challenges and have lower performances. To solve this problem, a number of analyses from the perspective of explainable AI (Grad-CAM++) have been conducted to explore which input region a CNN classifier looks at to distinguish a female from a male and vice versa, which helps to avoid extracting improper features that result in unsatisfactory performance. These analyses straightforwardly emphasize the importance of removing biases and increasing diversity for a training dataset, the guarantee of which will improve the generalization ability of the CNN model and thus render more fair results.

keywords: gender classification; thermal imagery; explainable CNN; Grad-CAM++; dataset biases

1 Introduction

Gender classification has been frequently used in many fields of our life, like access control, product recommendation, social statistics, etc. Regarding the popular trend of designing smart buildings, there is also a place for gender classification based on the notion that females and males usually have different thermal sensations in the same environment [1–8]. Therefore, smart buildings (specifically referring to a thermal adaptive architecture in this work that regulates its indoor microclimate according to individual thermal needs) can provide a more satisfactory microclimate for the occupants if the gender information of each occupant is taken into consideration.

Accordingly, automatic gender classification as a computer vision task has been more prevalent recently, as it can save a lot of manpower involved in the manual recognition of genders. Most of the gender classification works are targeted at visible facial images (see Fig. D.1) [9–15] and gait information of a whole body (see Fig. D.2) [16–20]. However, as to the application in a specific domain of a thermal adaptive built environment in real life, these studies are not that effective anymore. Because occupants are usually reluctant to have their facial information recorded by visible cameras for privacy concerns and it is much more difficult to extract gait information as sitting is the domi-

nant posture in such indoor environments. Therefore, gender classification in thermal imagery which is more privacy-friendly draws more attention in recent days.

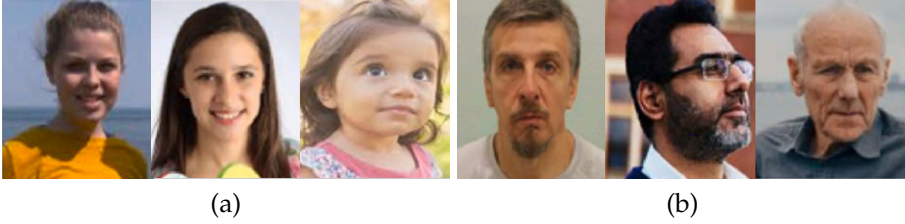


Fig. D.1: Example RGB images for gender classification. (a) Females. (b) Males. Image source: [9].



Fig. D.2: Human silhouettes in a gait cycle and the averaged gait energy image (the right one) for gender classification. Image source: [16].

As to gender classification for thermal inputs, researches are quite rare as expected since the number of thermal sensors is dramatically lower than that of visible sensors from our daily experiences. From these researches, some of them use thermal imagery as an auxiliary modality besides visible imagery to solve illumination problems caused by shadows or day and night shifts. Studies [21–23] are such cases that either use a late fusion (of classification scores of the thermal pipeline and the visible pipeline) or an early fusion (of the extracted thermal features and visible features) to get the final classification result. The difference is that [21, 22] use the full human body while [23] uses the facial image as the input for subsequent gender classification. Other works [24, 25] only use thermal imagery as the data source and correspondingly apply convolutional neural network (CNN)-based classifiers to predict the gender category for each acquired facial image.

Generally, these studies have demonstrated the potential of doing gender classification in thermal imagery. However, in these works, the CNN classifiers are working as a “black box” where only the input images, the output gender categories, and an accuracy rate are known. A lot more questions are unsolved. For example, why a classifier can grasp the ability to recognize a female from a male and vice versa; why some of the inputs are mis-classified while others are correctly-classified; how to improve the performance further,

etc. To answer these queries, a technology called explainable AI (artificial intelligence) can be employed to make the CNN processing pipeline more “transparent”.

Accordingly, this work uses explainable CNN to visualize the most discriminative regions that a classifier looks at to predict the gender of the person in the input image—the “where to look” problem. By doing a series of such visualization experiments in thermal imagery, we have found a severe influence of bias in the training set on the classification result and further emphasized the importance of a diverse dataset for the better generalization ability of a CNN model. Taking this into consideration (remove the bias and increase the diversity of the used dataset) can pave the way to a more robust and successful gender recognition in thermal imagery. As a whole, our contributions are:

- We have done extensive experiments on gender classification in thermal imagery and discovered that a CNN-based gender classifier focuses more on extrinsic features, especially the hair.
- Considering that extrinsic features are more prone to data imbalance/bias problems than intrinsic facial features, we have done ablation analyses to investigate how biases/stereotypes in a dataset influence the gender classification result and also recommended good practices to find biases more effectively in advance.

2 Explainable CNN

This work is characterized by using explainable CNN to investigate the gender-discriminative regions of an input image, especially in thermal imagery that has much fewer details for a CNN to use compared to visible imagery. Therefore, this section is centred on the explainable CNN which specifically refers to class activation mapping (CAM) [26]-based interpretation for a classification task.

2.1 Class Activation Mapping

For an image fed to a classical CNN composed of several convolutional layers, the first convolutional layer works on the image in a way of using multiple filters to do pixel-wise multiply and accumulate operations with the move of a sliding window; this processing generates a number of feature maps that will be fed to the subsequent layers where the same operations are repeated. Usually, the feature maps extracted in deeper layers have abstracted the input details into a general representation that can be used for semantic high-level computer vision tasks. Therefore, the feature maps of the last convolutional

layer are expected to have the best semantic information and also contain the spatial information before being converted to a fully-connected shape. Therefore, analyzing the influence of these feature maps on the prediction result (or any class) can localize the most important feature map region for this class, based on which the corresponding important region in the input is known.

In detail, if the feature maps of the last convolutional layer is a $H \times W \times C$ tensor (H , W , and C represent the size of height, the size of width, and the number of channels, respectively), we can imagine that each channel of these maps must have a different contribution to the predicted class (or any class). If this channel-wise contribution can be measured as a value, all the channels of feature maps can be integrated into a map of size $H \times W$ by a weighted summation. In this way, the value of each point location in the map can represent the importance of this location to the predicted class (or any class). Therefore, a series of subsequent procedures (including a visualization of this map based on the value of each point, an interpolation of this visualization to the input image size, and an overlapping of this interpolated map on the original input image) will generate a final class activation map that shows a class-discriminative location distribution as exemplified in Fig. D.3. In the Fig. D.3(b) and (c), the regions with red color contribute most to a certain class; in other words, the region of the dog head (or the cat head) exactly corresponds to the dog (or cat) class. Such a color distribution also gives a class activation map another name for convenience—heat map.

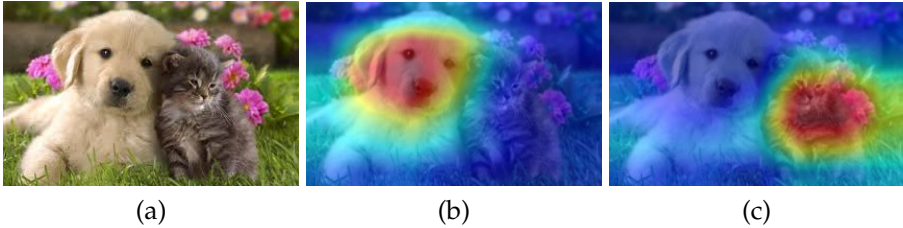


Fig. D.3: An example of class activation maps. (a) The input image. (b) The class activation map for the class of dog. (c) The class activation map for the class of cat. Image source: [27].

From this example, a CNN interpreted by CAM is much more “transparent” instead of a “black box”, explicitly explaining which region it looks at for a result so that we can get an understanding of “why it predicts what it predicts”. For this advantage, many works have used the technology of CAM in various applications. Work [28] uses CAM to visualize the heat maps of chest X-ray and CT-scan images regarding distinguishing COVID-19 cases from normal cases by a CNN classifier. These CAM visualizations point out the crucial locations of a lung image that should be paid more attention to and thus help medical workers give a faster and more accurate diagnosis.

Similarly, medical researches [29, 30] use the CAM technology to do diabetic retinopathy classification on fundus images and sclerosis types classification on brain MRI images, respectively. Both works have visualized the most discriminative image region for the diagnosis which helps to increase the diagnosis accuracy, reduce the examination time, understand the disease development, and improve the treatment. In other fields, CAM are applied to railway scene classification [31], iconography artwork analysis [32], examination of loosening bolts of automotive parts [33], etc.

2.2 Generating Class Activation Maps

As mentioned above, CAM has been applied to many applications. However, the exact research on gender classification in thermal imagery has not used it yet. Therefore, this specific task via CNN is still a “black box” where we do not know what happens and whether we can trust the result or not. To this end, visualizing the gender-discriminative region for an input is implemented in this work, for which each input’s class activation map has to be generated. Hence, this subsection explains how to generate it, specifically referring to Gradient-weighted Class Activation Mapping++ (Grad-CAM++) [34].

Grad-CAM++ is an improved version of Grad-CAM [35] which will be first introduced here. As mentioned in the last subsection, a class activation map is an interpolated version of the weighted summation of the feature maps from the last convolutional layers. Thus, calculating the weight of each feature map that represents its contribution to the predicted class (or any class) is the key. Inspired by the back-propagation in CNN, the gradient flowing from any class score to a certain neuron can represent its importance for the class. Intuitively, a large positive gradient means that the increase of the neuron value will result in a large increase in the class score. In contrast, a small positive gradient means that the increase of the neuron value only accounts for a limited increase in the class score. On the other hand, a negative gradient means the change of the neuron value and the change of the class score are in different directions.

From this understanding, the gradient of a class score y^c (for a certain class c) with respect to a pixel value A_{ij}^k at location (i, j) on the K th feature map A^k explicitly represents the location’s importance to y^c . Then according to equation D.1 [35], the gradients of all locations on A^k are globally averaged to a value α_k^c which represents the importance of A^k to y^c . This α_k^c is the weight of this channel A^k in calculating the class activation map. In the same way, the weights of other channels of feature maps are obtained. With them, the initial class activation map (heat map) considering all the channels by a weighted summation is obtained, in the form of D.2 [35] where a ReLU

operation filters out the influence of features that have a negative influence on the class score. After interpolating the initial heat map to the original input size and then overlapping it on the input, the final class activation map via Grad-CAM for the class c is obtained.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (\text{D.1})$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (\text{D.2})$$

This Grad-CAM scheme has been verified in many studies and successfully visualizes class-discriminative regions. However, the global average calculation in equation D.1 has sacrificed the spatial information, which may influence the localization ability to find the region of interest for a certain class. To solve this, Grad-CAM++ proposes that the gradient on each pixel location (i, j) on the K th channel A^k should have its own weight α_{ij}^{kc} (obtained via partial differential equation of higher order) when calculating the integrated weight w_k^c for A^k and thus improves equation D.1 to D.3 [34]. With other similar subsequent calculations, the class activation map of an input via Grad-CAM++ will be generated.

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot \text{ReLU}(\frac{\partial y^c}{\partial A_{ij}^k}) \quad (\text{D.3})$$

According to the comparisons in [34], Grad-CAM++ provides a more accurate localization ability in heat maps, and thus this work chooses Grad-CAM++ for all the analyses.

3 Explainable Gender Classification

This section gives the information of the used thermal dataset, the implemented CNN, the corresponding performances, and the detailed analyses of the results via Grad-CAM++.

3.1 Dataset and Implementation Information

Dataset Information

The thermal modality dataset from Tufts face database [36, 37] is chosen as the images to train a CNN gender classifier for the reasons: (1) faces are important regions for gender recognition not only from our daily experiences but also from existing works in RGB visible imagery; (2) according to the database description [38], the thermal modality dataset has recorded the

3. Explainable Gender Classification

faces of 112 participators including 37 females and 75 males, which makes itself a relatively large dataset to train a classifier. To be specific, this thermal dataset consists of two scenarios. One is the TD_IR_A (Tufts Dataset Infrared Around) subset captured by a FLIR camera which is moved to nine positions around each participator in an approximate semicircle shape; the other one is the TD_IR_E (Tufts Dataset Infrared Emotion) subset to record each person with five different emotions by the same FLIR camera. Example images of these two kinds of thermal subsets are shown in Fig. D.4 and Fig. D.5, respectively. The images from these two subsets have the same spatial resolution of 336×256 pixels. We, therefore, decide to combine them together for the gender classification task. Accordingly, the thermal images from all the 37 females and another 37 males constitute the training set. After an augmentation step of one horizontal flipping and two kinds of rotations (-5° and 5°), there are a total of 3108 female images and 3102 male images for training.

Instead of separating part of the images from the Tufts dataset into a validation set, we chose the thermal facial images from the VAP RGB-D-T dataset [39] for validation (see Fig. D.6). In detail, the validation set includes 40 images from all the eight females recorded in the VAP RGB-D-T dataset and another 40 images from eight males for a fair evaluation. It is worth mentioning that: (1) we have resized the validation images into the same size of 336×256 as the training set for convenience and better model generalization; (2) we do not prepare a testing set as the number of female participators is not enough for three subsets, and this work is primarily targeted at the visualized understanding of how a CNN does gender classification instead of only boosting the performance. As a whole, the information of the training set and validation set is listed in Table D.1.



Fig. D.4: Example images of the TD_IR_A (around) subset. Each person has nine images corresponding to nine directions. (a) Images of a female. (b) Images of a male.

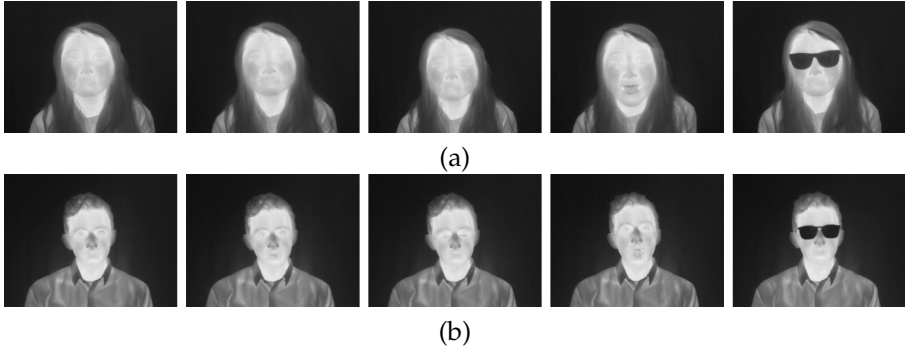


Fig. D.5: Example images of the TD_IR_E (emotion) subset. Each person has five images corresponding to five emotions: neutral, smile, closing eyes, shocked, with sunglasses. (a) Images of a female. (b) Images of a male.

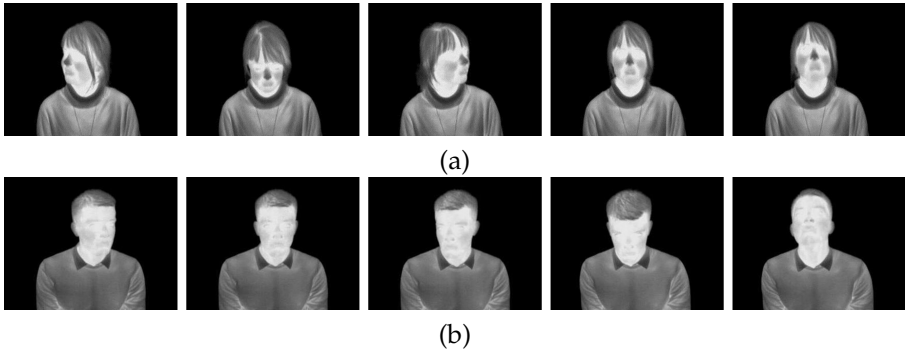


Fig. D.6: Example images of the validation set. Each person has five images corresponding to five directions. (a) Images of a female. (b) Images of a male.

Table D.1: The information of the training set and validation set for gender classification.

	Number of images		Number of participants	
	Training	Validation	Training	Validation
Female	518 augmented to 3108	40	37	8
Male	517 augmented to 3102	40	37	8

Implementation Information

To realize the CNN-based classification of genders, we finetuned AlexNet [40] whose last fully-connected layer (of a drop-out rate as 0.25) has two output neurons representing the predicted scores of the two genders. In detail, the AlexNet is finetuned for 50 epoches with the batch size of 64 and the Adam optimizer whose learning rate is $1e-4$, and then the model at the last epoch (at which the network is converged) is used to conduct the visualization analysis

by using Grad-CAM++ PyTorch toolbox [27]. It is also worth mentioning that we do not use very deep networks (more than ten convolutional layers) like VGG series [41], ResNet series [42], Inception series [43], etc., considering that the number of the thermal data is not large and the details of a thermal image are also fewer.

3.2 Results and Analyses

After 50 epochs, both the training and validation loss curves tend to be flat, indicating that the finetuned AlexNet has stabilized. And it achieves a classification rate of 100% on the validation set. The corresponding checkpoint is saved and used to analyze how the CNN does gender classification. Accordingly, Fig. D.7 shows some validation images and their class activation maps.

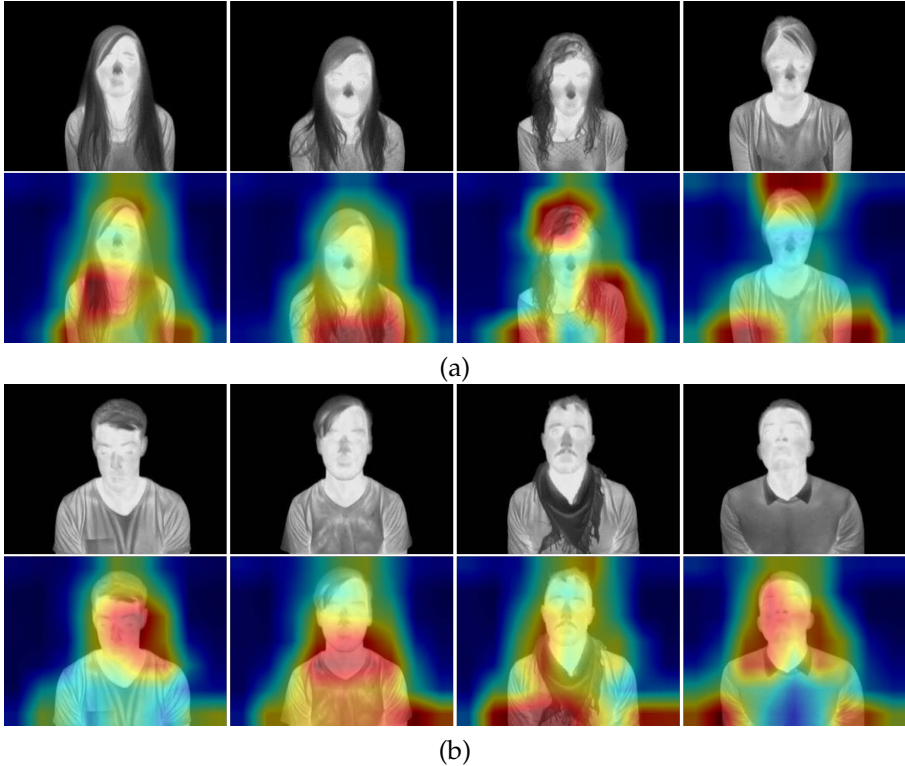


Fig. D.7: Some validation images and their class activation maps. (a) Females. (b) Males.

From Fig. D.7, it can be seen that the common discriminative regions are the hair area (for females), near the ear/neck/shoulder area (for males), or

near the upper part of the arms (for both genders). Based on our daily experiences, one hypothesis that females are expected to have long hair that can cover the ear, shoulder, neck, and upper arm regions is proposed to explain these highlighted areas. In a thermal image, the temperature of hair is lower than that of a bare-skin area but higher than the background temperature. Therefore, a hair area has a high potential to be recognized by the CNN, based on which gender classification is realized. To test the hypothesis, class activation maps of some training images are also generated and then shown in Fig. D.8.

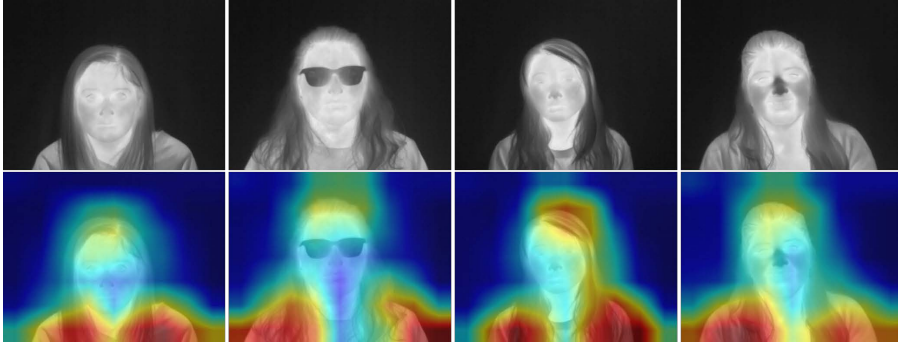


Fig. D.8: Some training images and their class activation maps.

From these figures, it is clear to see that a female's hair ends are always intensively activated, which indicates that the shoulder-length and longer hair (covering shoulders, necks, and part of the chest and upper arms) is an important sign of a female. Whereas the ear, neck, and shoulder region of a male usually have no signs of hair (as shown in Fig. D.7 (b)) that makes him distinguished from a female. Therefore, it is the hair area that the CNN looks at for gender classification on this dataset.

Extended Experiment: the Bias of Hairstyles

The above analysis points out the importance of hairstyles in gender classification. However, hairstyle is an extrinsic character that cannot intrinsically represent a person; besides, the CNN-learned knowledge of females having longer hair than males is a stereotype. Hence, a CNN looking at this region for gender classification is potentially dangerous as it is highly prone to biases.

The fact that the CNN learns such a stereotype certainly indicates that it has been impacted by the bias of hairstyles in the training phase. By manually examining the 37 females in the training set, we find out there are 19 females with shoulder-length or longer hair, two females with hijabs that are similar

3. Explainable Gender Classification

to long hair, 12 females tying a ponytail, and only four females with short hair above the ears. The number of long-hair females is almost five times that of females with short hair. Though this imbalance has not influenced the validation performance (due to the simplicity of the validation set), the learnt discriminative regions have been influenced by this hairstyle bias greatly. In other words, if we leave this bias alone, the trained CNN will always only look at the neck/shoulder/upper arms region to recognize a female or male and thus ignore other useful information, which will potentially lead to undesired classification results in a testing phase.

Therefore, a further investigation of how gender-discriminative regions react to this hairstyle bias is conducted, which includes two extended experiments: (1) there are only females with shoulder-length or longer hair and hijabs in the training set, noted as Exp1; (2) there are only females with tied ponytails and short hair in the training set, noted as Exp2. Both the extended experiments use the same validation set as the basic experiment for evaluation. Whereas the training sets are changed as listed in Table D.2.

Table D.2: The information of the training set and validation set for extended experiments to investigate the influence of hairstyle biases on gender classification.

		Number of images		Number of participants	
		Training	Validation	Training	Validation
Exp1	Female	294 augmented to 1764	40	21	8
	Male	294 augmented to 1764	40	21	8
Exp2	Female	224 augmented to 1344	40	16	8
	Male	224 augmented to 1344	40	16	8

Exp1 is an experiment to deliberately enhance the bias of hairstyle for gender classification. We, therefore, hypothesize that the learned features are stereotyped and unfair and thus degrade the performance. After the same training process as the basic experiment, the finetuned AlexNet only gets a classification rate of 80% on the same validation set. All the failed cases are those females with tied ponytails that are mis-classified as males. On the contrary, Exp2 is an experiment to remove the stereotype/bias of hairstyles in the training set so that the CNN can learn a series of fair and comprehensive features or the best intrinsic features to recognize a female or a male. Therefore, the expectation is that the new learned weight/model can pay attention to regions besides those shown in Fig. D.7 and Fig. D.8 for more robust performance. After the same training process, the model of the last epoch achieves a significantly-improved classification rate of 96.25% compared to Exp1. For the visualization-based analysis, the respective model of Exp1 and Exp2 is used to generate the class activation maps for the validation images that are mis-classified in Exp1, as shown in Fig. D.9.

From these figures, the activated areas generated by Exp2 (Fig. D.9(c)) are

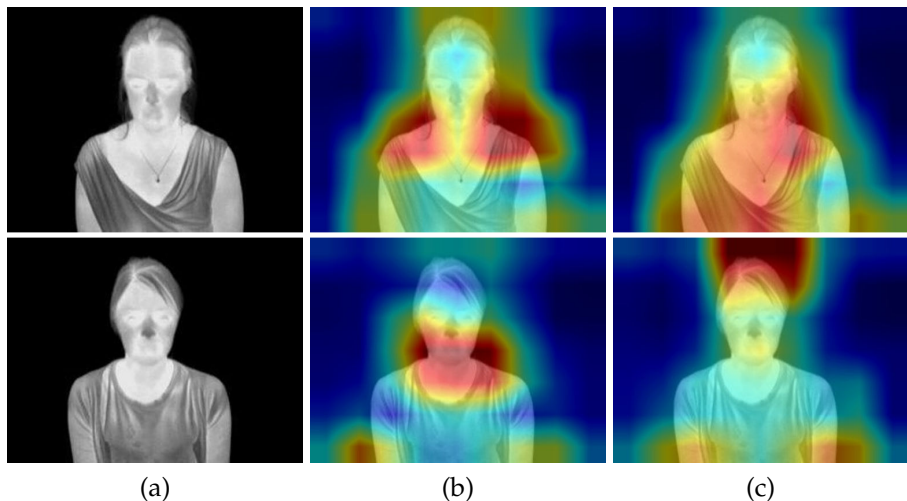


Fig. D.9: Class activation maps of validation images (mis-classified in Exp1). (a) The input images. (b) The class activation maps from Exp1. (c) The class activation maps from Exp2.

now either the top of the head or much broader compared to the neck areas (Fig. D.9(b)) that indicate shoulder-length hair in Exp1, making these failed cases in Exp1 correctly classified in Exp2. Therefore, removing the biggest bias of the hair length in the training set indeed helps the CNN to consider more aspects for a prediction.

Besides the mis-classified cases by Exp1, we have conducted a broader exploration for other validation images to compare the activated discriminative regions in Exp1 and Exp2. The procedures are: we first generated the five class activation maps of each validation participant’s five images but did not overlap these maps on their corresponding input images; then, for each person, we assembled the five maps into a single map from a channel/image-wise average; at last, we overlapped the single map on the frontal image of the corresponding person to get the final activation map. In this way, a final map grasps the ability to generally represent all the discriminative regions of an individual and provides a more comprehensive comparison between Exp1 and Exp2. For the illustration, these final maps are shown in the second column referring to Exp1 and the third column referring to Exp2 in Fig. D.10 and Fig. D.11. From these figures, the activated regions from Exp2 are always larger and more general than those from Exp1, demonstrating the more information the CNN looks at for a decision. Still, the hairstyle on the top of the head (Fig. D.10) is important, and the best intrinsic facial feature seems not to take a leading role from both experiments.

4. Discussion

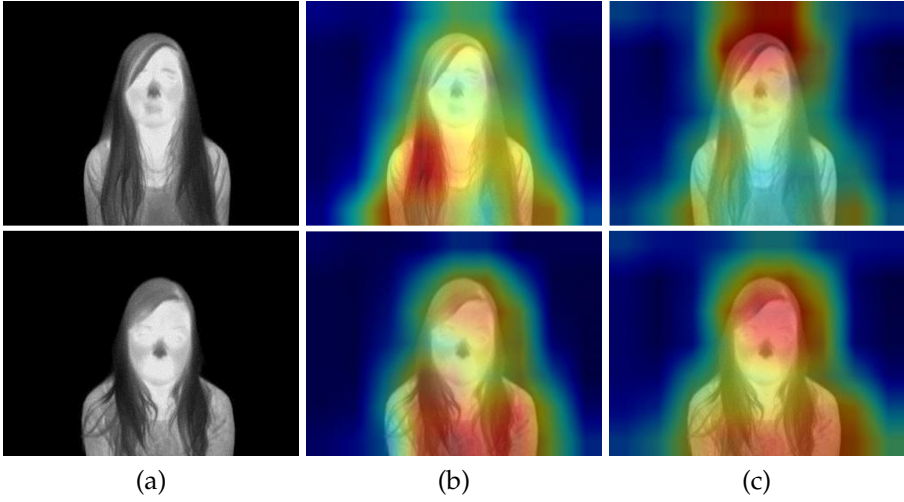


Fig. D.10: The assembled class activation maps of validation images. (a) The frontal image of a participant. (b) The assembled class activation maps from Exp1. (c) The assembled class activation maps from Exp2.

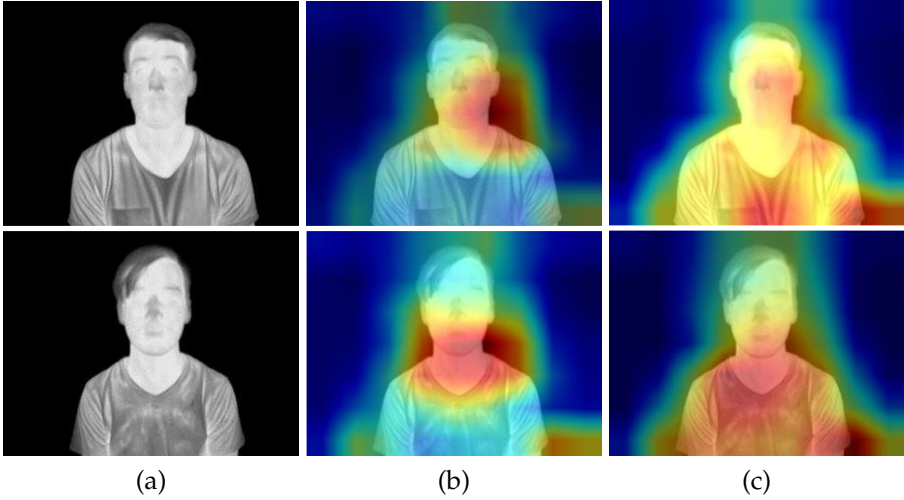


Fig. D.11: The assembled class activation maps of validation images. (a) The frontal image of a participant. (b) The assembled class activation maps from Exp1. (c) The assembled class activation maps from Exp2.

4 Discussion

From these experiments on thermal images of the Tufts dataset and the VAP RGB-D-T dataset, it is concluded that the CNN for gender classification in

thermal imagery focuses more on extrinsic features (especially the hair) instead of the intrinsic facial features that are normally used in identity recognition [44]. This is reasonable as more detailed information is lost in a thermal image compared to its visible counterpart.

However, extrinsic features are highly prone to data imbalance/bias problems. Hence, from the perspective of data collection for data-driven approaches, a large and diverse dataset, especially with no biases, is very important. On the one hand, any bias may cause a disastrous influence as the available thermal features are limited. On the other hand, finding biases may be much easier in thermal imagery as it only has a relationship with thermal radiations. Therefore, paying more attention to the object/phenomena with thermal/temperature properties is an effective way, for example, the time duration during which the data is collected as this may influence the environment temperature, any object that has a similar temperature distribution with humans (e.g. a warm-blooded animal) as it may be recognized as a person, any object that has a very different temperature distribution but has a relationship with humans (e.g. accessories like glasses as shown and explained in the next paragraph), etc. In contrast, much more stereotypes exist in visible imagery for gender classification, like the color of clothes, the makeup, the skin color, etc., which require more careful consideration.

Besides the above-mentioned recommendations to find biases in thermal imagery, a more practical method is to apply class activation mapping to the training images to see what the CNN has learnt, provided that the CNN has overfitted the training data so that it extracts the features of the training images completely. In this way, the activated regions on a training image will directly illustrate if there exists a data imbalance/bias. For example, a bias that only females wear glasses in the training set will make the corresponding glasses region activated as the class-discriminative area shown in Fig. D.12. And then, a male with glasses will be recognized as a female in the testing phase. This kind of bias usually has a severe impact on thermal datasets because the distinctive temperature difference between a person and his/her accessories (glasses) will enhance the influence of a nonliving thing on the performance.

5 Conclusion and Future Work

This work investigates the “where to look” problem regarding gender classification in thermal imagery, that is, to find the gender-discriminative region by using the technology of class activation mapping. On the basis of extensive experiments and corresponding analyses on two thermal datasets, we have discovered that a CNN-based gender classifier focuses more on extrinsic features. However, these extrinsic aspects are prone to introducing biases

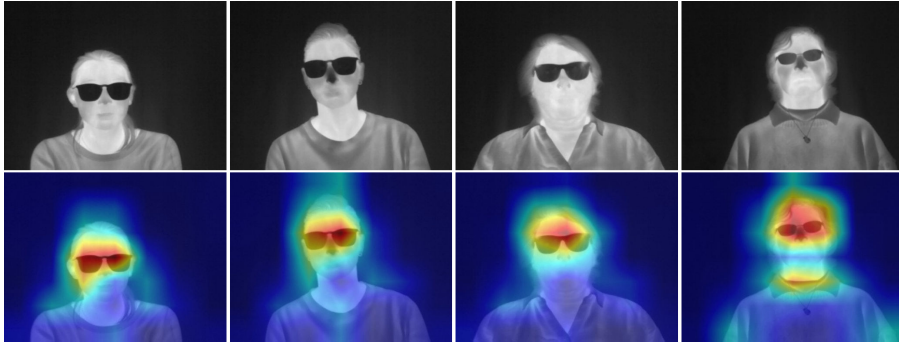


Fig. D.12: The class activation maps of some training images if there exists a bias that only females wear glasses in the training set. These images are from the Tufts dataset.

and stereotypes that will cause improper feature extraction and unfair results. Keeping an eye on this side-effect and then removing such biases in the training set can make the class-discriminative regions more general and reasonable, which has the potential to lead to more robust performance.

In the future, we plan to extend our experiments and analyses to more thermal datasets and scenarios to further testify our findings, for example, investigating the gender-discriminative regions for cropped face-only images. Besides, as we have emphasized the importance of the dataset diversity for improving performances, we plan to convert some benchmark RGB datasets to the thermal modality by using generative adversarial networks so that a lot more images/videos can be utilized for tasks in thermal imagery.

References

- [1] Sami Karjalainen, “Thermal comfort and gender: a literature review,” *Indoor air*, vol. 22, no. 2, pp. 96–109, 2012.
- [2] Jinhua Hu, Yingdong He, Xiaoli Hao, Nianping Li, Yuan Su, and Huaiddi Qu, “Optimal temperature ranges considering gender differences in thermal comfort, work performance, and sick building syndrome: A winter field study in university classrooms,” *Energy and Buildings*, vol. 254, pp. 111554, 2022.
- [3] Jéssica Kuntz Maykot, Ricardo Forgiarini Rupp, and Enedir Ghisi, “Assessment of gender on requirements for thermal comfort in office buildings located in the brazilian humid subtropical climate,” *Energy and Buildings*, vol. 158, pp. 1170–1183, 2018.

- [4] J  ssica Kuntz Maykot, Ricardo Forgiarini Rupp, and Enedir Ghisi, "A field study about gender and thermal comfort temperatures in office buildings," *Energy and Buildings*, vol. 178, pp. 254–264, 2018.
- [5] Madhavi Indraganti, "Gender differences in thermal comfort and satisfaction in offices in gcc and asia," in *Gulf conference on sustainable built environment*. Springer, 2020, pp. 483–497.
- [6] Thomas Parkinson, Stefano Schiavon, Richard de Dear, and Gail Brager, "Overcooling of offices reveals gender inequity in thermal comfort," *Scientific reports*, vol. 11, no. 1, pp. 1–7, 2021.
- [7] Mina Jowkar, Hom Bahadur Rijal, Azadeh Montazami, James Brusey, and Alenka Temeljotov-Salaj, "The influence of acclimatization, age and gender-related differences on thermal perception in university buildings: Case studies in scotland and england," *Building and Environment*, vol. 179, pp. 106933, 2020.
- [8] Madhavi Indraganti and Michael A Humphreys, "A comparative study of gender differences in thermal comfort and environmental satisfaction in air-conditioned offices in qatar, india, and japan," *Building and Environment*, vol. 206, pp. 108297, 2021.
- [9] James Rwigema, Joseph Mfitumukiza, and Kim Tae-Yong, "A hybrid approach of neural networks for age and gender classification through decision fusion," *Biomedical Signal Processing and Control*, vol. 66, pp. 102459, 2021.
- [10] Mingxing Duan, Kenli Li, Canqun Yang, and Keqin Li, "A hybrid deep learning cnn–elm for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, 2018.
- [11] Khalil Khan, Muhammad Attique, Ikram Syed, Ghulam Sarwar, Muhammad Abeer Irfan, and Rehan Ullah Khan, "A unified framework for head pose, age and gender classification through end-to-end face segmentation," *Entropy*, vol. 21, no. 7, pp. 647, 2019.
- [12] Sepidehsadat Hosseini, Seok Hee Lee, Hyuk Jin Kwon, Hyung Il Koo, and Nam Ik Cho, "Age and gender classification using wide convolutional neural network and gabor filter," in *2018 International Workshop on Advanced Image Technology (IWAIT)*. IEEE, 2018, pp. 1–3.
- [13] Ratinder Kaur Sangha and Preeti Rai, "An appearance-based gender classification using radon features," in *Data, Engineering and Applications*, pp. 159–169. Springer, 2019.

- [14] Juan E Tapia and Claudio A Perez, "Clusters of features using complementary information applied to gender classification from face images," *IEEE Access*, vol. 7, pp. 79374–79387, 2019.
- [15] Vidya Muthukumar, "Color-theoretic experiments to understand unequal gender classification accuracy from face images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.
- [16] Shiqi Yu, Tieniu Tan, Kaiqi Huang, Kui Jia, and Xinyu Wu, "A study on gait-based gender classification," *IEEE Transactions on image processing*, vol. 18, no. 8, pp. 1905–1910, 2009.
- [17] Jang-Hee Yoo, Doosung Hwang, and Mark S Nixon, "Gender classification in human gait using support vector machine," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2005, pp. 138–145.
- [18] Mohammed Hussein Ahmed and Azhin Tahir Sabir, "Human gender classification based on gait features using kinect sensor," in *2017 3rd IEEE International Conference on Cybernetics (Cybconf)*. IEEE, 2017, pp. 1–5.
- [19] Ankita Jain and Vivek Kanhangad, "Gender classification in smart-phones using gait information," *Expert Systems with Applications*, vol. 93, pp. 257–266, 2018.
- [20] Paola Barra, Carmen Bisogni, Michele Nappi, David Freire-Obregón, and Modesto Castrillón-Santana, "Gait analysis for gender classification in forensics," in *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Springer, 2019, pp. 180–190.
- [21] Dat Tien Nguyen and Kang Ryoung Park, "Body-based gender recognition using images from visible and thermal cameras," *Sensors*, vol. 16, no. 2, pp. 156, 2016.
- [22] Dat Tien Nguyen, Ki Wan Kim, Hyung Gil Hong, Ja Hyung Koo, Min Cheol Kim, and Kang Ryoung Park, "Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction," *Sensors*, vol. 17, no. 3, pp. 637, 2017.
- [23] Shangfei Wang, Zhen Gao, Shan He, Menghua He, and Qiang Ji, "Gender recognition from visible and thermal infrared facial images," *Multimedia Tools and Applications*, vol. 75, no. 14, pp. 8419–8442, 2016.
- [24] Muhammad Ali Farooq, Hossein Javidnia, and Peter Corcoran, "Performance estimation of the state-of-the-art convolution neural networks for

- thermal images-based gender classification system," *Journal of Electronic Imaging*, vol. 29, no. 6, pp. 063004, 2020.
- [25] Kateřina Přihodová and Jakub Jech, "Gender recognition using thermal images from uav," in *2021 International Conference on Information and Digital Technologies (IDT)*. IEEE, 2021, pp. 83–88.
- [26] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [27] Jacob Gildenblat and contributors, "Pytorch library for cam methods," <https://github.com/jacobgil/pytorch-grad-cam>, 2021, last accessed: April, 2022.
- [28] Harsh Panwar, PK Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, and Vaishnavi Singh, "A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images," *Chaos, Solitons & Fractals*, vol. 140, pp. 110190, 2020.
- [29] Hongyang Jiang, Jie Xu, Rongjie Shi, Kang Yang, Dongdong Zhang, Mengdi Gao, He Ma, and Wei Qian, "A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1560–1563.
- [30] Yunyan Zhang, Daphne Hong, Daniel McClement, Olayinka Oladosu, Glen Pridham, and Garth Slaney, "Grad-cam helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging," *Journal of Neuroscience Methods*, vol. 353, pp. 109098, 2021.
- [31] Bing Zhao, Ping Li, and MingRui Dai, "Visualization of railway scene classification model via grad-cam," in *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2018, pp. 281–286.
- [32] Nicolò Oreste Pincirolì Vago, Federico Milani, Piero Fraternali, and Riccardo da Silva Torres, "Comparing cam algorithms for the identification of salient image features in iconography artwork analysis," *Journal of Imaging*, vol. 7, no. 7, pp. 106, 2021.
- [33] Eunsol Noh and Seokmoo Hong, "Automatic screening of bolts with anti-loosening coating using grad-cam and transfer learning with deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 4, pp. 2029, 2022.

- [34] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.
- [35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [36] Sensing The Panetta Visualization and Simulation Research Laboratory, "The tufts face database," <http://tdface.ece.tufts.edu/>, last accessed: April, 2022.
- [37] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A Taylor, Arash Samani, et al., "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.
- [38] hpvisionlab, "Tufts-face-database," <https://github.com/kpvisionlab/Tufts-Face-Database>, last accessed: April, 2022.
- [39] Olegs Nikisins, Kamal Nasrollahi, Modris Greitans, and Thomas B Moeslund, "Rgb-dt based face recognition," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 1716–1721.
- [40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [41] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
- [42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [44] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, pp. 342, 2020.

References

Part III

Safe Harbor

Paper E

Supervised versus Self-supervised Assistant for Surveillance of Harbor Fronts

Jinsong Liu, Mark P. Philipsen, and Thomas B. Moeslund

The paper has been published in the
*Proceedings of the 16th International Joint Conference on Computer Vision,
Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 2021*

© 2021 SCITEPRESS

DOI: 10.5220/0010323906100617

The layout has been revised.

Abstract

Drowning in harbors and along waterfronts is a serious problem, worsened by the challenge of achieving timely rescue efforts. To address this problem, we propose a privacy-friendly assistant surveillance system for identifying potentially hazardous situations (human activities near the water's edge) in order to give early warning. This will allow lifeguards and first responders to react proactively with a basis in accurate information. In order to achieve this, we develop and compare two vision-based solutions. One is a supervised approach based on the popular object detection framework, which allows us to detect humans in a defined area near the water's edge. The other is a self-supervised approach where anomalies are detected based on the reconstruction error from an autoencoder. To best comply with privacy requirements both solutions rely on thermal imaging captured in an active harbor environment. With a dataset having both safe and risky scenes, the two solutions are evaluated and compared, showing that the detector-based method wins in terms of performances, while the autoencoder-based method has the benefit of not requiring expensive annotations.

keywords: safety; drowning; surveillance; thermal imaging; deep learning; human detection; anomaly detection

1 Introduction

More than 40 people drown every hour of every day. Drownings typically occur when children fall into ponds, pools or wells; passengers or workers fall overboard or sink with ships; as a consequence of floods or when people are drunk in the vicinity of water [1]. Clearly, the causes of drowning accidents are many, as are the solutions. Here, we specifically want to address the deaths that can be prevented in urban spaces and industrial areas that are associated with harbor fronts.

A drowning person must be rescued within a few minutes. Unfortunately, it takes around 6 minutes from the authorities being alerted till a rescue boat is in the water [2]. This means the chance of a successful rescue is greatly improved by early preparations and accurate knowledge of the person's position. This calls for a precautionary surveillance system to provide early warnings for hazardous situations in critical areas like Fig. E.1 shows.

Such a surveillance system is mostly fulfilled by manual video surveillance now. However, continuously monitoring large areas of waterfronts manually is inefficient. If the operators who monitor the streams can be assisted by an intelligent system, the efficiency will be much higher. Like a human, the assistant system should be able to understand what is safe vs. risky or normal vs. abnormal. In order to grasp the ability, the system must



Fig. E.1: Thermal surveillance imaging for detecting potentially dangerous situations and alerting authorities. An alert should be raised when someone crosses the red line.

rely on cues correlated with drowning accidents, among which the most important cue is human activity near the water's edge. Relying on this cue, we investigate two alternative solutions based on computer vision and deep learning:

- Supervised human detection: A person's location and thus distance to the harbor's edge is used to determine whether the surveillance operator should be notified.
- Self-supervised anomaly detection: Scenes near the harbor's edge are classified as either normal or abnormal using the reconstruction loss from an autoencoder. In our case we consider a scene with any human activity in it to be unsafe, which should be classified as an anomaly. This solution is based on the fact that human activity near the water's edge is very rare.

The contributions in this work can be summarized as: An assistant surveillance system realized by two practical solutions (supervised vs. self-supervised) is proposed to detect potential drowning accidents from harbor fronts. The two solutions are evaluated and analysed with respect to strengths and weaknesses.

2 Related Work

Although most methods, databases, and benchmarks for detecting people, activities, and anomalies differ in various ways from our harbor scenario, it is likely that many of their findings will be useful and can be transferred from RGB images to thermal images. For this reason we give an overview of related work.

2.1 Object Detection

With the advent of convolutional neural networks (CNN), object detection has grown rapidly. A modern detector usually consists of a backbone which is pre-trained on large databases like ImageNet [3], a neck composed of several top-down connects or down-top connects to reuse extracted features, and a head predicting the objects' class and bounding box coordinates. The effectiveness of many mainstream detectors such as Faster R-CNN [4], SSD [5], YOLO [6–10], and RetinaNet [11] has been proven in benchmarks such as MS COCO [12] and PASCAL VOC [13].

Besides working on these general object detection benchmarks, the detectors are applied to specific situations, like analyzing soccer matches by detecting players [14], detecting stalled vehicles from moving vehicles to prevent traffic accidents [15], detecting pedestrians in autonomous driving context [16], monitoring social distance by human detection to stop the spread of epidemics [17]. Note that the above-applied scenarios are all in RGB mode and the detectors' application to thermal mode remains underexplored.

2.2 Anomaly Detection

Anomalies are generally defined as incidents that are unusual and rare. This makes it difficult to gather a large balanced database to train a binary normal vs. abnormal classifier using supervised learning. Interestingly, with self-supervised learning the unbalanced nature of the problem can be turned into an advantage. For this reason, techniques such as autoencoders are popular for anomaly detection [18–23].

An autoencoder consists of an encoder and a decoder. The encoder learns to produce a compressed representation of the input, ending in a bottleneck. The bottleneck is the input to the decoder whose task is to reconstruct the original input from the compressed bottleneck representation. Both networks are trained by minimizing the difference between the input and its reconstruction. The core idea of self-supervised anomaly detection using autoencoders is to use normal data to train the autoencoder. This results in the autoencoder learning to faithfully reconstruct normal data while performing poorly with abnormal data. In this way, the reconstruction error can be used

to recognize anomalies, and the unbalanced nature of the data becomes an advantage.

This kind of methods for anomaly detection have been applied to datasets such as UCSD [24] and Avenue [25]. With the UCSD dataset, the aim is to classify the occurrence of carts, wheelchairs, skaters, and bikers as anomalies. With the Avenue dataset, anomalies include running and walking in the wrong direction as well as walking with bicycles. Again, note that these datasets are in RGB mode and the application of anomaly detection using autoencoders is unexplored when it comes to thermal mode datasets.

3 Challenges

In order to realize an assistant surveillance system for raising alarms to prevent drowning accidents, a range of challenges must be considered. These challenges include concerns such as privacy, challenges specific to thermal imaging, and a long tail of rare events.

3.1 Sensitive Data

According to the European general data protection regulation (GDPR) [26], personal data should be protected from being invaded and abused. To best comply with this set of rules, privacy-friendly thermal cameras are used, making it difficult to recognize a person in captured images.

3.2 Thermal Imaging

The use of thermal cameras is associated with benefits and drawbacks compared with RGB cameras. Thermal cameras can be used to capture people both day and night without the need of light sources. As thermal cameras rely on thermal radiation, temperature changes in the scene will influence the imaging. For instance, during warm days, the environment temperature will approach the temperature of the human body, resulting in a loss of contrast between the foreground (people) and the background.

The insulating properties of clothing also impact the appearance of people in thermal images, thus constituting a significant source of variation. Weather-induced phenomena such as wind, rain, and ice may also impact cameras installed outside. Moreover, the spatial resolution of thermal cameras is lower than visible light RGB cameras, which leads to the challenge of applying methods intended for high resolution RGB images to low resolution thermal images where the size of humans is relatively small.

3.3 Rare Phenomena



Fig. E.2: (a) Animal. (b) Reflection.

Rare and disturbing phenomena pose a challenge when developing an intelligent system since it is difficult to anticipate them. Fig. E.2 provides two examples from the harbor area: Fig. E.2(a) shows a red box around a dog which may be mistaken as a child; reflections due to water on the ground introduce false detections, as indicated by the two red boxes in Fig. E.2(b). Besides, as the same with other scenes, a person whose body is occluded severely or a person cluttered with a very similar background will make it difficult for any detector to work.

4 Applied Methods

As mentioned before, we believe both object detection and anomaly detection are worth pursuing for an assistant precaution system. This section will describe these two methods in detail. The approach based on object detection is illustrated in Fig. E.3(a). It processes frames individually as input and locate people in the image. If a detection is made on the water side of the red boundary, an alarm is raised. Fig. E.3(b) illustrates the autoencoder-based approach, where pixels from the water side of the red boundary are passed through the autoencoder and an alarm is raised if the input is poorly reconstructed, signifying an anomaly—human activity near the water’s edge.

4.1 Supervised Human Detection

To detect a human from a long distance a successful detector should have the ability to tackle small objects, and we value three aspects that matter to this

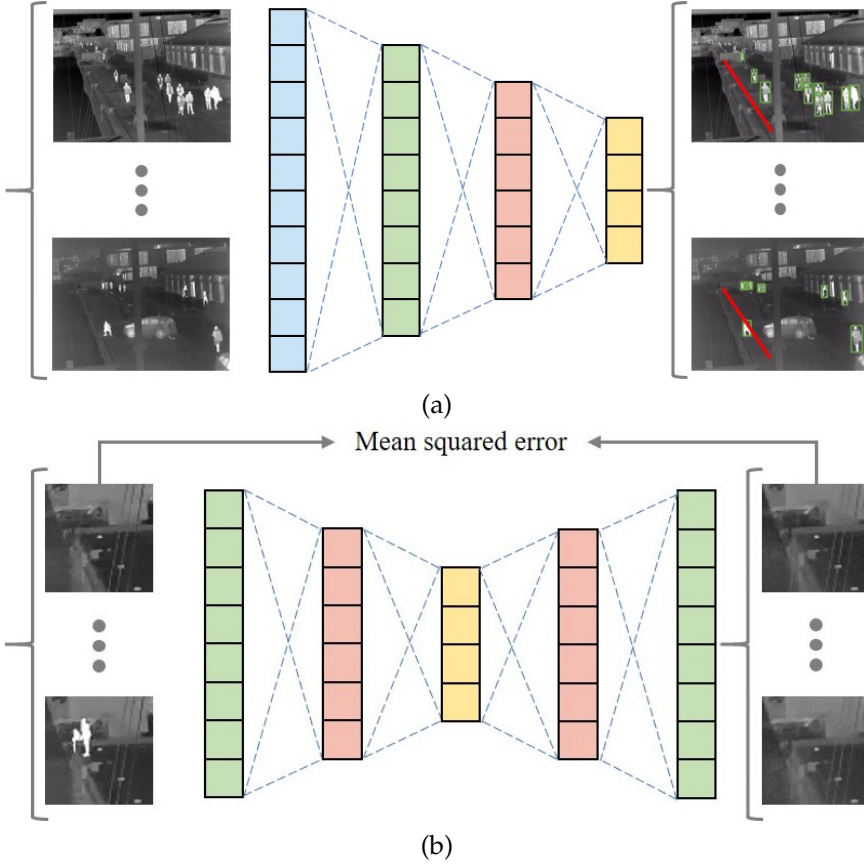


Fig. E.3: (a) Object detector where alerts are raised when people are detected on the risky side of the red boundary. (b) Autoencoder where the reconstruction error is used as an indicator of anomalies.

ability: anchor boxes, feature reuse, and scales, which are well designed in YOLOv5 [10]—the applied detector in the harbor scenario.

An anchor box gives the initial size of an object, and the predicted bounding box is the updated version of the anchor box that the object corresponds to. Therefore, the definition of anchor boxes is critical in a detector because an improper anchor box (either too large or too small) not only increases the prediction time but also leads to missing objects as this anchor box may have a very low intersection-over-union (IoU) with any ground truth box. For instance, to get a satisfactory performance on COCO database, YOLOv3 [8] uses a k-means clustering algorithm on COCO training set to define 9 anchors boxes, which emphasizes the importance of database-adaptive anchor boxes. That's why YOLOv5 is utilized here. Its capacity to dynamically de-

fine the number and sizes of anchor boxes according to the training set is of great benefit.

To accurately localize an object, appearance information from lower layers of a CNN is greatly helpful. But this information may vanish after passing through multiple layers in a deep network thus increasing the difficulty of object detection, especially for small objects. Feature reuse can address this problem by top-down or down-top bypass connections to combine features from both lower layers and deeper layers. It is to be noted that if the additional bypass itself has to go through deep layers, the efficiency of feature reuse will be reduced. YOLOv5 solves this reduced efficiency problem by introducing PANet [27] instead of FPN [28] as its network neck.

A detector with predictions at only one scale often fails for objects with different sizes. To address this issue, a detector should work on several scales, a way to avoid missing detections of small objects whose information may disappear in deeper layers. Therefore, small objects are detected with larger feature maps while large objects are detected with smaller feature maps. YOLOv5 predicts outputs on three scales which have different spatial resolutions, making it a good human detector for our task.

Safe vs. Risky Classification

If a person is detected, his/her relative location to the harbor's edge is the key to determine whether an alarm should be raised. Therefore, an alarm region near the water is predefined empirically. In Fig. E.4, the red line represents the alarm boundary expressed by Equation E.1, and the points $p1 = (67, 180)$ and $p2 = (170, 23)$ are the two endpoints of the line segment. For a person detected in the xy coordinate system of the image, if any coordinate (x_p, y_p) within the bounding box results in a z_p in Equation E.2 smaller than 0, the person is deemed inside the alarm region.

$$1.53x + y - 283 = 0 \tag{E.1}$$

$$z_p = 1.53x_p + y_p - 283 \begin{cases} \geq 0, & \text{safe} \\ < 0, & \text{risky} \end{cases} \tag{E.2}$$

4.2 Self-supervised Anomaly Detection

In order to measure human activities in regions near the water, we train an autoencoder formed by a standard 9-layer CNN structure, where the 5-layer encoder and 5-layer decoder share a bottleneck having 8 channels. In the encoder the convolutional filters increase in numbers (32, 64, 128, 256) while the feature maps decrease in sizes along with layers going deeper. Inversely,



Fig. E.4: Alarm region defined as the area between the red line and the water.

the decoder consists of transposed convolutions where the filters decrease in numbers (256, 128, 64, 32) and feature maps increase in sizes as the network approaches the output. With exception of the penultimate layer of the decoder, the output of each layer is batch-normalized and uses a LeakyReLU activation function. And the autoencoder is trained using normal images consisting of the background with the Mean Squared Error (MSE) as its loss function¹.

Reconstruction Error as a Measure of Activity

In order to explore the usefulness of reconstruction error as a measure of human activity, we use a square region (of size 64×64) near the water's edge. From this region 35,809 thermal images with very limited or no human activities are collected across hours and then used as normal frames to train an autoencoder. After training, a continuous sequence of 1250 thermal images, with significant human activities in some frames, is fed into the autoencoder to explore the change in reconstruction error over time.

This exploration is illustrated in Fig. E.5. The four images in Fig. E.5(a) are samples from the 4 locations represented by the red lines in Fig. E.5(c), and Fig. E.5(b) shows the reconstructions by the autoencoder for their corresponding inputs in Fig. E.5(a). From Fig. E.5(b), it is clear that the autoen-

¹Code at <https://github.com/JinsongCV/Supervised-Versus-Self-supervised-Assistant-for-Surveillance-of-Harbor-Fronts>.

4. Applied Methods

coder fails to reconstruct humans. This results in a high reconstruction error represented by the blue graph in Fig. E.5(c). As more people enter the area, the error increases further. This exploration proves the potential for monitoring human activity and detecting anomalies by self-supervised learning with the reconstruction error as a measurement.

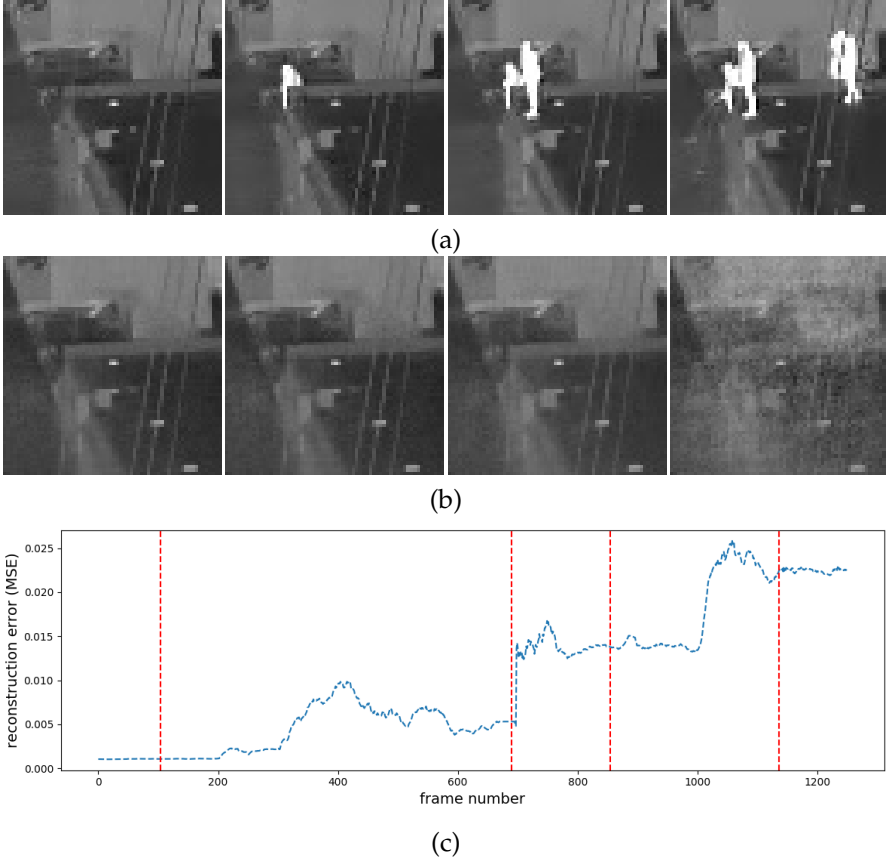


Fig. E.5: (a) Input patches from the region near the water's edge. (b) Corresponding reconstructions from the autoencoder. (c) The blue graph illustrates the reconstruction loss as across the sequence and the vertical red lines correspond to the samples shown in (a) in the same order.

Normal vs. Abnormal Classification

A threshold, as defined in Equation E.3, can be used to classify the data as normal (safe) or abnormal (risky). It can be determined either manually or automatically. If determined manually, the threshold will be based on the tacit knowledge and experience of the human operator. An automatic

threshold can be determined from a small labeled dataset, where the threshold reaching the desired balance between precision and recall for anomaly alarms is chosen.

$$\begin{cases} MSE < threshold, & normal(safe) \\ MSE \geq threshold, & abnormal(risky) \end{cases} \quad (E.3)$$

To enable comparison with the detector-based method, an alarm region is defined between the red line segment defined in Section 4.1 and the edge of the harbor (see Fig. E.6). This region is transformed to a rectangle (of size 64×192) as the input to the autoencoder by using OpenCV's warpPerspective function.



Fig. E.6: Region for anomaly detection by the autoencoder.

5 Experiments

This section gives the dataset information and experiments to prove the feasibility of both solutions.

The thermal camera is placed below the bridge to cover a popular walking path. To evaluate the two methods, the videos (of size 384×288) from February 3, 2020 to March 3, 2020 were collected by an authorized computer which protects the data from invasions. Different types of weather (rainy days, snowy days, windy days, and sunny days) occurred during this period, making the database more diverse and less biased. To consider the challenges

5. Experiments

of contrast, weather, and rare phenomena mentioned before, we manually selected and annotated 2358 images, which were then divided into the training set (1715), validation set (143), and test set (500). Note that this manual annotation means labeling bounding boxes for human detection. Besides, to fairly compare both solutions, autoencoder-based anomaly detection also uses the same 2358 images for training and evaluation. The experimental platform consists of a machine equipped with a NVIDIA GeForce RTX 2080Ti, Ubuntu 16.04 LTS, CUDA 9.2, Python 3.7.0, and PyTorch 1.6.0.

5.1 Supervised Surveillance Assistant

YOLOv5s [10] is fine-tuned by stochastic gradient descent with momentum 0.9 from a pretrained model on COCO dataset. The learning rate is set as 0.001. The training phase stops at 120th epoch where the network has converged. Other settings remain the same with the original YOLOv5s.



Fig. E.7: A failed case which should have raised an alarm. The red box refers to the undetected person.

In the testing phase, the best model on the validation set is used to do detection on the test set, achieving an average precision (AP_{50}) of 97.70%. Besides AP_{50} , the accuracy of true alarms for risky situations is also measured. Among the 500 test images, 91 of them have persons existing in the alarm region defined in Section 4.1. Based on the human locations predicted by YOLOv5s and Equation E.2, 85 out of the 91 images are classified as risky situations and no false alarms are raised, indicating a recall of 93.41% and a precision of 100%. All the 6 failed cases are related to undetected persons

having very small sizes. One example is shown in Fig. E.7 where the red box refers to the undetected person. As YOLOv5s is applied to frames from videos, it is likely that the undetected person will be detected in the earlier or later frames. As a whole, no matter with AP_{50} or with alarm rates, the human detection-based method works well.

5.2 Self-supervised Surveillance Assistant

In order to produce comparable results to the supervised surveillance assistant, the alarm region defined in Fig. E.6 is cropped and transformed from the same training set and test set used for the human detector. The autoencoder is trained from scratch for 200 epochs using the Adam optimizer with a learning rate of 0.0005. The experiments regarding anomaly detection using an autoencoder includes: (1) Automatically determining a threshold. (2) Investigating the sensitivity to abnormal data in the training set.

Finding a Suitable Threshold

As mentioned in Section 4.2, a threshold for the reconstruction error can either be determined manually or automatically. Here, we suggest computing the threshold by optimizing the F1 score on the training dataset (including 1628 normal images and 87 abnormal images). With the maximal F1 score of 0.917, this leads to a threshold at 0.000597 MSE.

Sensitivity to Abnormal Training Data

It is labor-intensive to make sure that the training set contains only normal patterns. For this reason we want to investigate the sensitivity of the method to small amounts of abnormal data. We compare training with two datasets, one consists of 1628 normal patterns, and the other consists of 1715 in total (1628 normal and 87 abnormal images). Table E.1 shows the two versions' performances on the test set, expressed as the area under the precision-recall curve (AUC). The slightly lower performance with the inclusion of abnormal images demonstrates the method's sensitivity to abnormal training data and supports the suspected conclusion that the occurrence of abnormal data is detrimental in the training set.

Table E.1: Performance comparison between two models trained on different datasets. "Normal" is trained on 1628 normal images. "Normal+abnormal" is trained on a set containing an additional 87 abnormal images.

Model	AUC
Normal	0.929
Normal+abnormal	0.904

5. Experiments

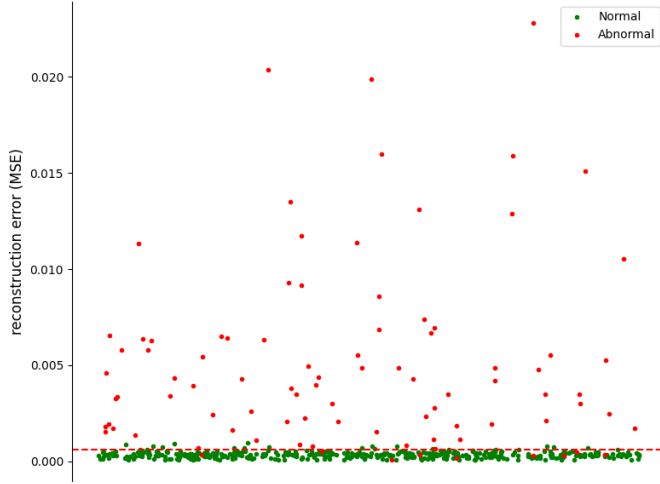


Fig. E.8: Distribution of normal (green dot) and abnormal (red dot) samples along with the decision threshold (red line). 21 False Positives (FP), 9 False Negatives (FN), 82 True Positives (TP), 388 True Negatives (TN).

Distribution of Normal and Abnormal Samples

With a threshold at the MSE of 0.000597, the best performing model achieves a recall of 90.11% and a precision of 79.61% on the test set. Fig. E.8 shows the distribution of normal and abnormal samples in the test set along with the threshold found using the F1 score. We expect normal frames (green dot) to generally be placed underneath the red line and abnormal frames (red dot) to be placed above the line.

To further investigate the failures in Fig. E.8, 4 cases are selected (two correctly and two wrongly classified) shown in Fig. E.9. Specifically, Fig. E.9(a) is a normal image mis-classified as abnormal due to the higher heat absorption and reflection of the harbor’s concretes and metals; Fig. E.9(c) is an abnormal image mis-classified as normal because of the low signal of human activity as a person is just entering the scene from the right side. This indicates the challenge of providing a reasonable standard of anomaly especially when a person has just entered the region. As a contrast, Fig. E.9(b) and Fig. E.9(d) are classified correctly.

This person-entering problem originates from a simple automatic annotation based on human locations. Any coordinate (within a ground truth bounding box of a person) located in the alarm region results in a labeling as abnormal. To reduce the unfair influence of such entering phenomena, we manually sort both the 1715 training set and the 500 test set, resulting in an additional “entering” category that will be disregarded. Therefore, the train-

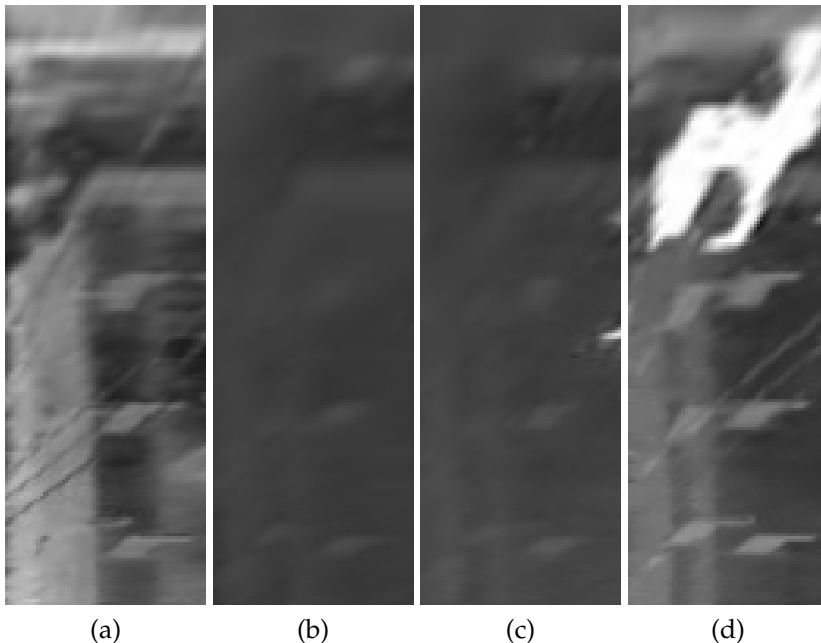


Fig. E.9: (a) A normal image with loss of 0.00095. (b) A normal image with loss of 0.00006. (c) An abnormal image with loss of 0.00029. (d) An abnormal image with loss of 0.02281.

ing set is now composed of 1628 normal images and 30 abnormal images, without 57 entering images; the new test set is composed of 409 normal images and 79 abnormal images, without 12 entering images. The experiment mentioned in Table E.1 is redone with the new datasets, and the results can be seen in Table E.2, where “Normal” and “Normal+abnormal” are the same models from Table E.1. Because the entering images are removed, the AUC is much better on the reduced test set (409+79). “Normal+clean abnormal” means that the model is trained on the new training set (1628+30) without entering images.

Table E.2: Performance comparison on the test set having no entering images. “Normal” and “Normal+abnormal” correspond to the models in Table E.1. “Normal+ clean abnormal” is trained on the training set without entering images.

Model	AUC
Normal	0.995
Normal+abnormal	0.974
Normal+clean abnormal	0.975

6 Discussion

Failure Modes The object detector is prone to FN due to unconventional appearance, occlusion, and clutter. The autoencoder on the other hand benefits from unconventional appearance but suffers from FP due to unusual backgrounds such as higher heat reflections.

Training Effort The object detector requires a large number of annotations. If this can be achieved, a detector can perform well in the vast majority of scenes without additional fine-tuning or reconfiguration. The autoencoder on the other hand requires retraining for each scene. In return, it requires no labeling or very limited labeling, which means it can be adapted to a specific problem with little effort.

Future Work In the future we want to consider temporal information and depth information for better differentiation of activities and image homography to remove perspective influences. Besides, the two approaches will have to be evaluated using a much larger and more diverse dataset to ensure that these solutions are workable all year across multiple locations.

7 Conclusion

We compare two alternative vision-based methods for assisting the surveillance of harbor fronts with a high risk of drowning accidents. One method utilizes object detection to detect people in low resolution thermal images and to raise warnings when people are detected inside a risky area. The detector is able to perform this task with perfect precision and a high recall of 93.41%. It fails in situations with occlusion and clutter. The other method uses an autoencoder and measures human activity based on the reconstruction error between input frames and the autoencoder's reconstructions. The autoencoder-based approach achieves a recall of 90.11% and a precision of 79.61%. It fails due to unusual background phenomena such as heat reflections and people only partially entering the monitored region. Given that the two methods have different strengths and weaknesses, one or the other might be more appropriate depending on the application.

References

- [1] WHO, "Global report on drowning: preventing a leading killer," <https://www.who.int/news-room/fact-sheets/detail/drowning>, 2014, last accessed: September, 2020.
- [2] , " <https://tryghavn.create.aau.dk>, last accessed: September, 2020.

- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multi-box detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [7] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.
- [8] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [9] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [10] Ultralytics, "Yolov5," <https://github.com/ultralytics/yolov5>, 2020, last accessed: October, 2020.
- [11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

- [14] Pier Luigi Mazzeo, Paolo Spagnolo, Marco Leo, and Tiziana D’Orazio, “Visual players detection and tracking in soccer matches,” in *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2008, pp. 326–333.
- [15] Linu Shine, Anitha Edison, and Jiji C. V. , “A comparative study of faster r-cnn models for anomaly detection in 2019 ai city challenge,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [16] Zhixin Guo, Wenzhi Liao, Yifan Xiao, Peter Veelaert, and Wilfried Philips, “Deep learning fusion of rgb and depth images for pedestrian detection,” in *30th British Machine Vision Conference*, 2019, pp. 1–13.
- [17] Narinder Singh Punj, Sanjay Kumar Sonbhadra, and Sonali Agarwal, “Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deepsort techniques,” *arXiv preprint arXiv:2005.01385*, 2020.
- [18] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis, “Learning temporal regularity in video sequences,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.
- [19] Yong Shean Chong and Yong Haur Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” in *International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.
- [20] Trong-Nguyen Nguyen and Jean Meunier, “Anomaly detection in video sequence with appearance-motion correspondence,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1273–1283.
- [21] Elvan Duman and Osman Ayhan Erdem, “Anomaly detection in videos using optical flow and convolutional autoencoder,” *IEEE Access*, vol. 7, pp. 183914–183923, 2019.
- [22] Hao Song, Che Sun, Xinxiao Wu, Mei Chen, and Yunde Jia, “Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos,” *IEEE Transactions on Multimedia*, 2019.
- [23] K Deepak, S Chandrakala, and C Krishna Mohan, “Residual spatiotemporal autoencoder for unsupervised video anomaly detection,” *Signal, Image and Video Processing*, pp. 1–8, 2020.
- [24] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, “Anomaly detection in crowded scenes,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1975–1981.

References

- [25] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.
- [26] Paul Voigt and Axel Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed.*, Cham: Springer International Publishing, 2017.
- [27] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.
- [28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

Paper F

Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift

Ivan Nikolov, Mark P. Philipsen, Jinsong Liu, Jacob V. Dueholm,
Anders S. Johansen, Kamal Nasrollahi, and Thomas B.
Moeslund

The paper has been published in the
*Proceedings of the Neural Information Processing Systems (NeurIPS) Track on
Datasets and Benchmarks, 2021*

© 2021 Individual authors and Neural Information Processing Systems Foundation Inc.

The layout has been revised.

Abstract

The time dimension of datasets and long-term performance of machine learning models have received little attention. With extended deployments in the wild, models are bound to encounter novel scenarios and concept drift that cannot be accounted for during development and training. In order for long-term patterns and cycles to appear in datasets, the datasets must cover long periods of time. Since this is rarely the case, it is difficult to explore how computer vision algorithms cope with changes in data distribution occurring across long-term cycles such as seasons. Video surveillance is an application area clearly affected by concept drift. For this reason we publish the Long-term Thermal Drift (LTD) dataset. LTD consists of thermal surveillance imaging from a single location across 8 months. Along with thermal images we provide relevant metadata such as weather, the day/night cycle and scene activity. In this paper we use the metadata for in-depth analysis of the causal and correlational relationships between environmental variables and the performance of selected computer vision algorithms used for anomaly and object detection. Long-term performance is shown to be most correlated with temperature, humidity, the day/night cycle and scene activity level. This suggests that the coverage of these variables should be prioritised when building datasets for similar applications. As a baseline, we propose to mitigate the impact of concept drift by first detecting points in time where drift occurs. At this point we collect additional data that is used to retraining the models. This improves later performance by an average of 25% across all tested algorithms.

1 Introduction

Once computer vision algorithms step outside the lab and are deployed in real-life outdoor applications, their performance tends to drop significantly due to conditions changing over time, i.e. concept drift [1–3]. Concept drift can materialize as gradual, recurring or sudden changes in the visual representation of the scene. Existing datasets, in general, favour coverage of multiple locations [4, 5] for short periods of time [6–8]. Such datasets are ill suited for exploring long-term effects such as concept drift and algorithms developed on their basis are unlikely to show robustness to long-term phenomena. Research studying concept drift [9, 10], uses synthetic datasets or datasets augmented in order to introduce drift. This does not necessarily completely represent real-world concept drift.

Our work presents a novel real-world dataset covering the 8 months from January to August. This time span means that the dataset encompasses a wide range of weather conditions, human activity, seasonal transitions, and recurring cycles such as weekdays, weekends, mornings and evenings. Along with the thermal images, timestamped metadata has been gathered. The

metadata includes weather data such as temperature, humidity, precipitation, etc. as well as metrics for scene activity level. We use the dataset to study concept drift by exploring contributing factors and demonstrating their effects on algorithmic performance. By publishing the dataset, we seek to aid the community in evaluating exiting algorithms against a long-term benchmark and in the development of algorithms that show greater robustness to long-term phenomena.

To explore the dataset, two common tasks are chosen, namely anomaly and people detection. These tasks tend to suffer strong performance degradation when exposed to long-term concept drift [11]. Object detection in general or detecting people in particular is a fundamental task involved in many use cases such as autonomous driving [12–14], tracking [15–18] and re-identification [19–21]. Common for many of the use cases is the application of object detection in unconstrained environments and across long spans of time. Anomaly detection, where the goal is to detect unusual behavioral patterns, is another task that is exposed to concept drift. These algorithms must be able to distinguish irrelevant changes due to e.g. concept drift from emergencies such as burglaries or assaults [5], car accidents [22], loitering and suspicious behaviour [23], indoor [24] and outdoor [25–27] falls.

We select representative algorithms for each task and evaluate their performance across time and in relation to environmental factors. As expected, all models exhibit performance degradation, as the test data diverges from the training set. Temperature and humidity prove to influence the models the most, followed by the change between day and night and the activity level of the scene. On the other hand, variation in precipitation and wind do not influence the performance of the models. In general, methods that learn from solving tasks that consider the entirety of the image are likely to be less impacted by drift, compared to methods that consider small regions or individual pixels [28]. An example could be object detectors vs. autoencoders, where something like brightness is likely to impact the autoencoder’s reconstruction significantly, but won’t effect the class or position of objects. By including both autoencoders and object detectors we ensure that both ends of this spectrum are covered in our analysis.

Finally, a baseline algorithm is presented to reduce the consequences of concept drift. This algorithm provides additional training data from points in time where concept drift is detected. This baseline is intended to encourage researchers to develop other methods of reducing the impact of concept drift. We believe that our findings on this novel dataset generalize to other environments and use cases, as well as other modalities and therefore will be an example to follow for future definition and collection of datasets. This in turn will help the community getting closer to deploying long-term computer vision algorithms for real-life outdoor applications. The main contributions of this paper can be summarized as follows:

- The Long-term Thermal Drift (LTD) dataset—the longest-spanning systematically collected thermal dataset comprised of 8 months of video data, containing both timestamp and weather condition metadata;
- In-depth analysis of the correlational and causal relationships between the performance of models and environmental factors;
- A baseline algorithm for reducing the effects of concept drift.

2 Related Work

2.1 Concept Drift Detection

As many systems need to be deployed and work stably for long periods of time and with input data which can change both gradually and suddenly, the presence of drift and ways to deal with it is a topic that has been widely studied. In computer vision it is normally studied by either focusing on specific real-world use cases or synthetically augmenting existing datasets. Real-world cases can be taken from egocentric video [29] or industrial inspection [30]. These cases present both examples of the problem and detection methods, but have limited use outside of the specific environments. Augmented versions of popular datasets such as MNIST and CIFAR can also be used. The works by [10] and [31] focus on methods for detecting data shifts using differences between the training and testing data, utilizing dimensionality reduction and statistical tests like Maximum Mean Discrepancy and Kolmogorov-Smirnov test. The benefit of using synthetically augmented data for testing is that different types of shifts can easily be simulated—from gradual drift to adversarial attacks [9]. But these simulated shifts do not always correspond to real-world ones. Some more robust methods also exist [11], aimed at using real-world drift in wider variety of use cases. The need for more research into concept drift, paired with a long-term real-world dataset is evident, as the effects from it can limit long term deployment of vision systems [32, 33].

2.2 Datasets

We can separate previous work roughly in two types of use cases—datasets that contain a scene from a stationary location, like the ones captured from CCTV and surveillance cameras and datasets with constantly changing locations, like the ones specifically directed towards autonomous cars, robots and human egocentric footage. The two types of datasets are used for different tasks, like vehicle and pedestrian detection and environmental segmentation

for changing datasets [4, 34, 35] and pedestrian tracking and anomaly detection for stationary ones [7, 36, 37]. The changing datasets also benefit from more diverse data coming from different sensors, compared to more image based stationary datasets. Our proposed LTD dataset is directed towards advancing the state-of-the-art in stationary location outdoor urban datasets by providing a longer duration, larger variation and rich metadata. A comparison in Table F.1 shows how the dataset stacks up against previous work.

Datasets used for autonomous driving with changing locations [34, 35, 40, 41] contain multiple modalities like LiDARs, RGB, depth cameras, as well as GPS and IMU data. They also contain data with longer duration from multiple days [43] to a whole year [39]. These datasets also focus on presenting adverse weather conditions, which can be used for domain adaptation and making autonomous driving and robotics application more robust [35, 42, 43]. Thermal datasets are less prevalent but still widely used [38, 49]. These moving location car datasets normally do not contain explicit information of their duration, as they are captured from many cars and the data is sampled.

On the other hand stationary location datasets do not contain any information about the period over which they were collected. This combined with the relative short duration of many of the widely used datasets ([7, 44, 45, 50]) makes it impossible for them to be used for studying long-term effects on deployed machine learning solutions. The duration of some of these datasets is taken from the research presented in [47]. Some larger datasets are gathered from internet videos [5], which lack the needed continuity for testing gradual concept drift in the data. More recent datasets have been produced with the goal to capture larger variations in the environments [37, 47], but with a limited scope. The lack of metadata is another problem, limiting the study of factors causing concept drift, as only some of the investigated datasets provide insufficient metadata [37, 48, 51]. Most of the investigated datasets focus on RGB data, with only some containing both RGB and thermal data [4, 37]. However, thermal imaging is better at preserving people’s anonymity as it does not capture facial and body details. This removes the need for post-processing like blurring or pixelating faces to protect personal data [52–54], which is a crucial requirement for complying with the European general data protection regulations (GDPR). The thermal imaging market has seen significant growth [55] and is forecast to expand even more in the following years [56, 57], which makes it necessary for long-term public thermal datasets to be easily accessible.

3 The Long-term Thermal Drift (LTD) Dataset

To address the gaps seen in the stationary surveillance state-of-the-art and to leverage the need for more thermal data, a new dataset is proposed. It

3. The Long-term Thermal Drift (LTD) Dataset

Table E1: Existing urban computer vision stationary and changing location datasets. The *Location* can be either changing denoting moving camera like the ones on self-driving cars or stationary like on surveillance cameras. The *Type* of the datasets can be either RGB, thermal or LiDAR. The *Duration* is the size of the dataset in hours. The *Period* is the capturing time span and the *Metadata* is any additional information.

Name	Year	Location	Type	Duration [hours]	Period	Metadata
KAIST [4]	2015	Changing	RGB/Thermal	43.41	-	-
CVC-14 [38]	2016	Changing	RGB/Thermal	11.8	-	-
Oxford RobotCar [39]	2017	Changing	RGB/LiDAR	-	1 year	GPS, IMU, Day/Night, Weather
Aachen Day-Night [40]	2018	Changing	RGB	-	-	GPS, Day/Night, Weather
Gated2Depth [41]	2019	Changing	RGB/LiDAR	-	-	GPS, IMU, Day/Night, Weather
Dark Zurich [42]	2019	Changing	RGB	-	-	GPS, Day/Night
ACDC [43]	2020	Changing	RGB	-	several days	GPS, Weather
Ford AV [35]	2020	Changing	RGB/LiDAR	-	1 year	GPS, IMU Day/Night, Weather, Time
Bdd100k [34]	2020	Changing	RGB	-	-	Weather, Time
UCSD [44]	2010	Stationary	RGB	3.1	-	-
Caltech Pedestrian [45]	2011	Stationary	RGB	10	-	-
VIRAT [46]	2011	Stationary	RGB	29	-	-
Avenue [6]	2013	Stationary	RGB	0.5	-	-
ShanghaiTech Campus [7]	2018	Stationary	RGB	3.6	-	-
Surveillance Videos [5]	2018	Stationary	RGB	128	-	-
Street Scene [36]	2020	Stationary	RGB	4	2 summers	-
ADOC [47]	2020	Stationary	RGB	24	1 day	-
AU-AIR [48]	2020	Stationary	RGB	2	-	Time, Positions
MEVA [37]	2021	Stationary	RGB/Thermal	144	3 weeks	GPS, Time
LTD (Our)	2021	Stationary	Thermal	298	8 months	GPS, Day/Night, Weather, Time

consists of thermal videos with resolution 288×384 captured through the period of **8 months** using a Hikvision DS-2TD2235D-25/50 thermal camera [58]. The camera is a long wavelength infrared (LWIR) unit, capturing wavelengths between 8 and $14 \mu\text{m}$. Raw data is captured through the day and saved in a mp4 format as 8-bit uncalibrated grayscale videos. A pre-processing algorithm is then run through the data. It first cuts the raw files into days starting from 00 : 00 and separates them into folders. Each folder is timestamped with the year, month and day timestamp. The videos for each day are then cut into **2-minute** clips selected from every 30 minutes through the day, for a total of **298 hours**. These videos are additionally timestamped with hour and minute timestamp. The starting point of the data is May 2020 until September 2020, together with a second part from January 2021, up until May 2021. This gives the data a large weather variation through the winter, spring and summer seasons. The images were taken on the harbor front in Aalborg, Denmark. The approximate longitude and latitude coordinates are given as (9.9217, 57.0488). We provide the dataset—<https://www.kaggle.com/ivannikolov/longterm-thermal-drift-dataset>, together with the code to extract the necessary data and to reproduce the experimental pipeline <https://github.com/IvanNik17/Seasonal-Changes-in-Thermal-Surveillance-Imaging>.

Some examples of seasonal and day and night variation of the captured data, together with weather and human activity variation can be seen in Fig. F.1. These large variations, together with a total size almost twice as large as other datasets in Section 2.2, allow for studying the effects of concept drift on trained models.

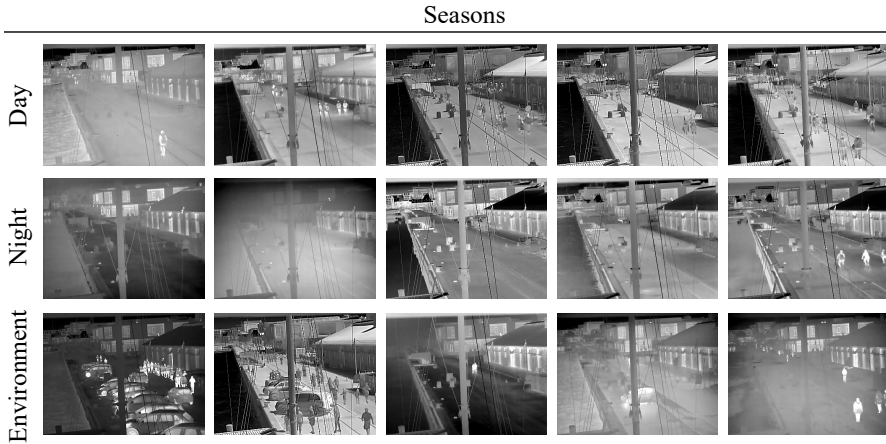


Fig. F.1: Examples of extreme changes in the image data contained in the proposed dataset. From left to right the day and night rows show example changes from data of February, March, April, June and August. The third row shows changes based on weather conditions and human activity.

Fig. F.1 depicts issues stemming from the natural thermal data concept drift, such as grayscale inversion in the background and people in different seasons, view limitation and reflections caused by weather like fog, rain, snow, view cluttering from multiple people and vehicles.

3.1 Metadata Analysis

Besides video data we also provide metadata in the form of weather data, gathered using the open source Danish Meteorological Institute (DMI) weather API [59] in 10-minute intervals. The selected properties are—temperature, measured in $^{\circ}\text{C}$, relative humidity percentage measured 2m over terrain, accumulated precipitation in $[\text{kg}/\text{m}^2]$, dew point temperature in $^{\circ}\text{C}$ measured 2m over terrain, wind direction in degrees orientation, wind speed in $[\text{m}/\text{s}]$, both measured 10m over terrain, mean sun radiation in $[\text{W}/\text{m}^2]$ and minutes of sunshine in the measured interval. These properties are selected, as it is speculated that they would be useful to explain changes in the captured image data. An overview of the average weather metadata measurements of the dataset can be seen in Table F.2. Temperature and relative humidity have been shown to affect thermal cameras, when detecting surface defects in concrete structures [60], measuring skin temperature changes on athletes [61], getting accurate readings for volcanology [62] and inspecting food [63]. Precipitation and dew point temperature can indicate the presence of rain, fog or high moisture and condensation. These can increase attenuation of infrared light and change the produced camera response [64, 65]. The build-up of moisture can create puddles in the images, which would change the scene reflectivity and reflected temperature [66]. The sun radiation and amount of sunshine can affect the captured images by rapidly changing the intensity of the infrared light. Finally wind speed and direction can cause movement of background parts of the scene like water ripples, ropes, etc., as well as movement of the camera itself.

Table F.2: Average metadata for each month. From left—temperature, humidity, precipitation, dew point, wind direction, wind speed, sun radiation and minutes of sunshine in a 10-minute interval.

	Temp. [$^{\circ}\text{C}$]	Hum. [%]	Precip. [kg/m^2]	Dew P. [$^{\circ}\text{C}$]	Wind Dir. [degrees]	Wind Sp. [m/s]	Sun Rad. [W/m^2]	Sun [min]
Jan.	-0.48	90.10	0.01	-1.96	161.91	2.58	23.97	0.90
Feb.	-0.54	85.15	0.01	-2.83	131.00	2.95	51.12	1.42
Mar.	3.75	83.61	0.01	0.93	218.80	3.58	99.35	1.85
Apr.	4.47	97.25	0.13	4.10	126.50	2.97	67.31	2.23
May	10.74	75.46	0.01	6.07	217.32	3.04	256.76	3.66
June	16.36	71.46	0.01	10.57	151.27	2.37	256.46	3.63
July	12.91	75.32	0.01	8.46	268.15	3.97	270.17	3.62
Aug.	16.93	79.17	0.02	12.69	163.18	2.08	197.86	3.15

4 Long-term Performance Experiment

We study the effects of concept drift on six machine learning models—two autoencoders, two object detectors and anomaly detectors. For these experiments only weather parameters not found to have significant correlation to other parameters are considered, namely—temperature, humidity, wind speed, wind direction and precipitation. More information on the correlation between weather parameters is given in the Appendix.

4.1 Data Selection Protocol

In order to keep the experiments and labelling effort manageable, samples across the full data set are selected based on the following protocol. This is done to minimize the number of frames and maximize the variation covered by the selection. For the sampling temperature metadata is used, as it is proven to directly correlate with changes to thermal images [60, 61, 63]. The protocol can be summarized as follows:

1. Every **2-minute** clip in the dataset is sampled with a frequency of one frame per second, resulting in **120 frames per clip**;
2. Based on the temperature metadata, we select a cold month for the training set and another cold month, a median temperature one, and a warm month for the test set;
3. The training set exists in three variants: coldest day 13th of February, the corresponding week 13-20 of February, and the entirety of February;
4. The test sets consist of data from January (similar cold month), April (month with median temperature), and August (warmest month).

From each of the thus created subsets, a greedy furthest point sampling is used for selecting frames. The frames for each day are sampled by calculating the farthest distances in the 2D feature space of the frame numbering and the temperature. A visual example of the sampling can be seen in the Appendix. The amounts of selected samples vary for the training data depending on the used algorithm. This is further discussed in the next sections.

4.2 Tested Models

Six deep learning models are tested. All six are originally designed to work with RGB data, so their input channel is reduced from 3 to 1, corresponding to a change to the grayscale thermal data. No additional changes were made, as the focus of the paper is not algorithm performance but change in performance over time.

Two of the tested models are autoencoders, as representatives for dimensional reduction, noise removal, concept drift detection and anomaly detection methods. Autoencoders are well suited for researching concept drift in long-term datasets, as their reconstruction performance is inherently tightly connected to the training data. The first autoencoder follows a simple fully convolutional architecture with symmetric 5-layer encoder and decoder. The implementation is based on the autoencoder used in a previous work [27]. It is theorized that its simplicity will make it sensitive to concept drift in the input data. The second autoencoder is the latest version of the Vector Quantised Variational Autoencoder (VQVAE2) [67]. This autoencoder uses collections of multi-scale hierarchical discrete tensors, called codebooks, to map its latent space. This gives it more robustness compared to regular autoencoders. The VQVAE2 implementation used here is closely based on [68]. Both autoencoders are trained for 200 epochs.

Two versions of the anomaly detector method MNAD [69] are also tested. They extend traditional autoencoders, by introducing memory-guided normality detection. We look at the typical reconstruction based version (MNAD_recon), as well as the prediction approach (MNAD_pred) using the preceding four consecutive frames to predict the future frame. The backbone consists of the U-Net structure, without skip-connections for the MNAD_recon variant. In between the encoder and decoder of U-Net is a memory module, storing prototypical events, concatenated with the original encoder output. The memory is primarily learned during training, but also updates during testing. Both versions are trained for 100 epochs.

Lastly two supervised object detectors are also tested—the YOLOv5 and Faster R-CNN [70]. The chosen hyperparameters for YOLOv5 remain the same as the work in [71], except that the initial learning rate is set to 0.00075 and trained for 200 epochs. The Faster R-CNN is trained for 200 epochs as well with SGD, with initial learning rate set as 0.005, the weight decay as 0.005 and the momentum kept at 0.9. Both object detectors have previously been successfully applied to outdoor thermal imaging [72–75].

The autoencoders are trained on a NVIDIA GTX1070 Super, the anomaly detectors on a NVIDIA RTX3080 and the object detectors on a NVIDIA RTX2080Ti.

4.3 Drift Algorithmic Performance Analysis

This experiment aims to see how the performance of the selected algorithms changes depending on the variation of the training data.

The training sets for the autoencoders and the anomaly detectors contain 5000 frames per subset, sampled using the method discussed in Section 4.1, where 20% are used for validation. Performance is reported as the average MSE across every image in each of the three test sets. The performance of the

two autoencoders and anomaly detectors is listed in Table F.3. We can see that the MSE for the CAE, VQVAE2 and MNAD_recon increases the farther away the test data goes from the training data. It can also be seen that the larger temporal pool provided for sampling for the weekly and monthly training data helps with keeping the MSE lower through the different months. The MNAD_pred is the only model keeping a consistent performance through the months without any noticeable drift. This is most likely due to the U-Net skip connections being able to reconstruct the background scene with a very low reconstruction error.

Table F.3: Results are reported as the average of the MSE across every frame in the test set. Higher results show worse performance.

Methods	Train	Test		
	Feb.	Jan.	Apr.	Aug.
CAE	Day 5k	0.0096	0.0202	0.0242
	Week 5k	0.0061	0.0167	0.0212
	Month 5k	0.0042	0.0109	0.0147
VQVAE2	Day 5k	0.0051	0.0072	0.0068
	Week 5k	0.0039	0.0066	0.0061
	Month 5k	0.0021	0.0039	0.0035
MNAD Recon.	Day 5k	0.0028	0.0057	0.0069
	Week 5k	0.0065	0.0066	0.0062
	Month 5k	0.0015	0.0041	0.0048
MNAD Pred.	Day 5k	0.0008	0.0007	0.0009
	Week 5k	0.0007	0.0006	0.0007
	Month 5k	0.0007	0.0006	0.0007

For the object detectors, because of the necessary data-labeling a smaller number of images are used for training and testing—both having 100 frames per subset. In addition to these a validation set comprising of 51 images evenly sampled from a previous annotated dataset [27] collected in February 2020 is used. All of the subsets are annotated with bounding boxes around people seen in each frame using the LabelImg open source program [76]. The annotations are also part of the LTD dataset. Since the performance of object detector is based on detected bounding boxes, mAP is used to evaluate it. The performance of the object detectors is given in Table F.4. The accuracy of both object detectors, drastically drops in the month of April. To prevent overfitting the smaller amount of training data, we also observe the validation and test loss.

As a conclusion from the performance analysis the higher variation provided by sampling from the week and month data, has been translated to better and more stable models in all the tested models. We can still see the effects of the seasonal drift, so additional analysis will be provided in the

5. Drift Analysis

Table F4: Results are reported as the mAP_{50} across every frame in the test set. Lower results show worse performance.

Methods	Train	Test		
	Feb.	Jan.	Apr.	Aug.
YOLOv5	Day 100	0.8010	0.5390	0.5240
	Week 100	0.7940	0.4540	0.4860
	Month 100	0.7930	0.4860	0.4830
Faster R-CNN	Day 100	0.6760	0.3230	0.3370
	Week 100	0.6740	0.2790	0.3060
	Month 100	0.6400	0.2560	0.3180

following sections.

5 Drift Analysis

In this section we look at the possible relations between the observed model performance drift and the changes in the captured metadata. Looking through the data examples given in Fig. F.1, two main visual change types are identified—seasonal and day/night. These types can be caused by either changes in the weather conditions, the human activity or a combination between the two. The relation between the model performance metrics and metadata features representing these changes is analysed. As discussed in Section 3.1, we choose temperature, humidity, precipitation, wind direction and wind speed as weather data features. For analysing the day/night changes the timestamp data is used to calculate hours of the day, as well as to calculate the sunrise and sunset times [77, 78]. To quantify the activity in the scene the difference between each testing frame and the previous frame is calculated. The mean value from this difference is selected. To focus only on the scene activity, everything in the background that moves like the waterfront and the visible ropes and masts is masked out. More information on this can be found in the Appendix.

We choose to use the results only from the models trained on the monthly February data, for easier visualization. The correlation between each of these features and the measured performance metric for each of the methods is first calculated. For the autoencoders and anomaly detectors this is the MSE, while for the object detectors we calculate the F1-score from all images containing people, as it gives a good overview of the precision and recall of the models. Both the basic Pearson’s correlation, as well as Distance correlation that is more sensitive to non-linear relations are calculated [79, 80]. The statistical significance p-values are also calculated with threshold at 0.05. The calculated correlation r values are given in Table F.5 where those with

p-values below the threshold are shown in red.

From Table F.5 it can be seen that temperature and humidity have the largest correlation values to most of the metrics, as well as the most consistent statistically significant results, followed by the scene activity and day/night features. We focus on these four features in the following analysis.

To get a better understanding of not only the correlational, but also causal relations between the models' performance metrics and the chosen features, we look at the Granger causality test [81]. The test only guarantees a predictive causality between variables, but would be able to point out any possible connections. The Granger causality tests the null hypothesis that the past values of one variable do not cause another. The p-value threshold is set to 0.05, below that the null hypothesis can be rejected, with the conclusion that there is a predictive causality between the variables. As the normal Granger causality test as presented in [82] is used on data with linear relations, we also use the more robust non-linear Neural Granger test [83]. Two best performing versions are used, based on long-short term memory networks (LSTM) and multi-level perceptron (MLP). Both models were trained using proximal gradient descent [84], with $\lambda = 0.002$, ridge regression coefficient 0.01 and learning rate of 0.005. The results from the Granger causality tests are given in Table F.6, where cells shown with green indicate a statistically significant presence of Granger causality and the ones with red—no presence.

The results show that the human activity has no predictive causality towards the performance of the models, which combined with the results from the correlation analysis, can point towards a second-hand relation. Our hypothesis is that the change in weather conditions and the day/night cycle are related to the change in human activity. From the other features, temperature and the day/night cycle have stronger predictive causality towards the autoencoders and anomaly detectors, while humidity has a more balanced predictive causality.

Fig. F.2 shows the relationship between the features and the model metrics. As a processing step before plotting the temperature and humidity they are first smoothed using a mean filter with a kernel size of 20 and then the MSE is normalized between 0 and 1. This is done as they cannot be directly compared, but the trend of their change can be visualized. We plot the average values for the training month of February, as a vertical red line, to indicate a "threshold".

6 Drift Prediction Baseline

As a baseline for exploring and mitigating the effects of concept drift a reference algorithm for predicting drift is presented. We use three strongest features—temperature, humidity and day/night cycle, together with MSE

Table E5: Correlation between the model's measured performance values MSE and F1-score and the weather, time and scene activity features. Two correlation measures are used—Pearson's (P.C.) and Distance (D.C.) correlation. Measures which do not meet the statistical significance threshold of their p-values are shown in red and marked ✗. The Day/Night features is specified as D./N.

	Measure	Temp.	Hum.	Wind Dir.	Wind Sp.	Precip.	Activ.	D./N.	Hour
CAE - MSE	P. C.	0.679	0.636	0.018 ✗	0.157	0.109 ✗	0.270	0.545	0.166
	D. C.	0.682	0.588	0.158	0.170	0.126 ✗	0.291	0.538	0.287
VQVAE2 - MSE	P. C.	0.381	0.690	0.001 ✗	0.194	0.172	0.217	0.403	0.124
	D. C.	0.347	0.639	0.174	0.201	0.224	0.217	0.382	0.213
MNAD Recon. - MSE	P. C.	0.607	0.672	0.016 ✗	0.173	0.126	0.220	0.509	0.156
	D. C.	0.617	0.629	0.188	0.177	0.155	0.252	0.501	0.273
MNAD Pred. - MSE	P. C.	0.107 ✗	0.277	0.064 ✗	0.152	0.072 ✗	0.677	0.369	0.137
	D. C.	0.231	0.348	0.154	0.172	0.086 ✗	0.665	0.462	0.312
YOLOv5 - F1-score	P. C.	0.261	0.258	0.102 ✗	0.011 ✗	0.096 ✗	0.124 ✗	0.047 ✗	0.009 ✗
	D. C.	0.293	0.283	0.146 ✗	0.094 ✗	0.135 ✗	0.255	0.113 ✗	0.174 ✗
Faster R-CNN - F1-score	P. C.	0.354	0.456	0.115 ✗	0.135 ✗	0.0124 ✗	0.199	0.147	0.001 ✗
	D. C.	0.334	0.460	0.228	0.149 ✗	0.065 ✗	0.231	0.163	0.118 ✗

Table E6: Results from calculating linear and non-linear (LSTM and MLP) Granger causality tests. The cells marked with ✓ show positive predictive causality, while cells marked with ✗ show no significant causality.

	Temp.			Hum.			Activ.			D./N.	
	Basic	LSTM	MLP	Basic	LSTM	MLP	Basic	LSTM	MLP	Basic	MLP
CAE - MSE	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓
VQVAE2 - MSE	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓
MNAD Recon. - MSE	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓
MNAD Pred. - MSE	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✓
YOLOv5 - F1-score	✓	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗
Faster R-CNN - F1-score	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓

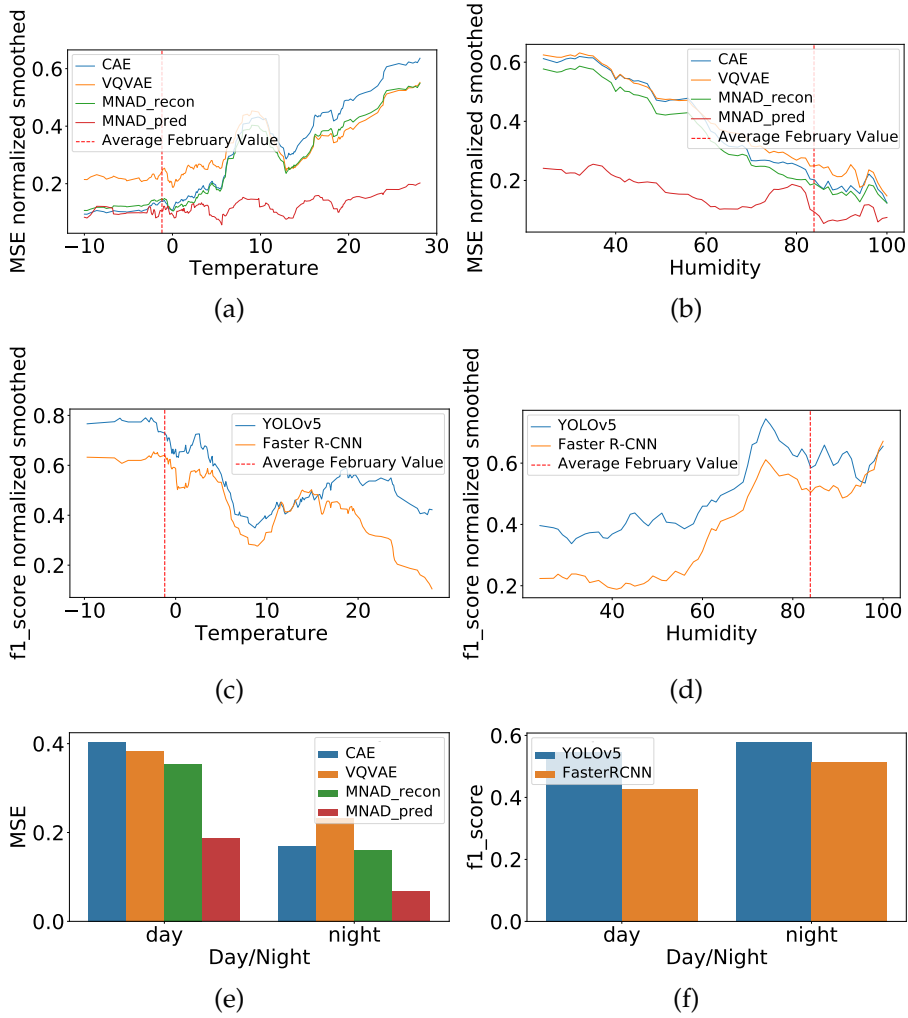


Fig. F2: Visual representation of the changes of MSE and F1-score for the tested models compared to the temperature, humidity and day/night cycle.

from our convolutional autoencoder (CAE) trained on the February monthly data. The CAE is chosen, as it is the most sensitive to changes in the dataset and is strongly correlated to the performance of all other tested models, except Faster R-CNN. The CAE MSE results from the training data are used together with the chosen features to train two widely used novelty/outlier detection models— isolation forests [85] and one-class SVM [86], available as part of scikit-learn [87]. The isolation forest has 100 base estimators, and the one-class SVM has a radial basis function (RBF) kernel and γ of 0.03. We

then test the results from each day from the full LTD dataset to detect points where many outliers emerge in both predictors. The first large concentration of outliers in 7 consecutive days is selected, which in our case is 5th of March.

To test if taking into consideration the data from the found drift point can help with the performance of the models against concept drift, training data from one week starting after the 5th of March is sampled. The new data is used together with the previous training data from February to retrain the tested models. The results, together with the month results from Table F.3 and F.4 for comparison, are given in Table F.7 and Table F.8. By adding the March data, all tested models achieve better results. We can see that the outlier detection models trained on the CAE MSE, together with the temperature, humidity and day/night cycle can be used together as an indicator for the amount of drift present in the input data.

Table F.7: The MSE results from the full month in Table F.3, compared to the ones using the new training datasets containing a combination of February and the week in March where drift is detected. Higher results show worse performance.

Methods	Train	Test		
		Jan.	Apr.	Aug.
VQVAE2	Feb. 5k	0.0021	0.0039	0.0035
	Feb. 5k + Mar. 5k	0.0020	0.0033	0.0030
MNAD Recon.	Feb. 5k	0.0015	0.0041	0.0048
	Feb. 5k + Mar. 5k	0.0006	0.0015	0.0025
MNAD Pred.	Feb. 5k	0.0007	0.0006	0.0007
	Feb. 5k + Mar. 5k	0.0007	0.0005	0.0006

Table F.8: The mAP₅₀ results from the full month in Table F.4, compared to the ones using the new training datasets containing a combination of February and the week in March where drift is detected. Lower results show worse performance.

Methods	Train	Test		
		Jan.	Apr.	Aug.
YOLOv5	Feb. 100	0.7930	0.4860	0.4830
	Feb. 100 + Mar. 100	0.8690	0.6640	0.6110
Faster R-CNN	Feb. 100	0.6400	0.2560	0.3180
	Feb. 100 + Mar. 100	0.6990	0.3910	0.3380

7 Conclusion and Future Work

In this paper we introduced the Long-term Thermal Drift (LTD) dataset spanning 8 months for detecting concept drift in deep learning models. The

dataset and the accompanying metadata can be used to document performance degradation as data drifts from the training set. These effects were studied on anomaly and object detection models, as well as autoencoders. It was demonstrated that more diverse training data lowers the effects of concept drift. The performance of the models showed a strong correlational and causal relationship to the change in temperature and humidity. A less pronounced relationship was observed to the day/night cycle and scene activity. Lastly, we showed how the concept drift can be further mitigated by detecting when it starts to manifest and providing additional data to the training process.

The proposed LTD dataset contains a combination of diverse environmental images and granular metadata. The equally spaced long-term data can be used to test the change in performance of deep learning models at different data scenarios—only day or night data, changes between activity in the week-day and weekends, summer and winter scenarios. The influence of weather conditions like rain, snow or fog can also be explored. The possibility of training more robust models and predicting when steps need to be taken before their performance degrades, is only possible with such long-term sequential datasets.

Possible negative social impacts of such long-term datasets concentrating on a single location is that they can be used to track the habits, interactions and movements of people. We offset this by providing a thermal dataset, which provides greater protection of people’s anonymity than conventional RGB imagery and does not require post-processing for blurring facial features.

The long-term nature of the dataset can also be used, as demonstrated in this paper, to do time-series analysis procedures on the outputs from different layers of deep learning models, from simple time-series analysis and forecasting models like Vector Autoregressive (VAR) Models [88] to more complex and data agnostic models like STRIPE [89] or Adversarial Sparse Transformers [90].

We believe that the proposed dataset and the accompanied analysis would help researchers understand the causes for performance drift in models and hence enable easier deployment of long-term solutions in outdoor environments.

8 Appendix

8.1 Metadata Correlation

The see all possible correlations between the captured weather data, the Pearson correlation matrix is calculated in Fig. F.3. The p-values for all correla-

8. Appendix

tions are close to 0, making them statistically significant. We additionally set a threshold of significant correlation above 50%. We can divide the weather data roughly in three categories: (1) correlated to temperature: dew point (correlation of 0.85) and sun radiation intensity (correlation of 0.54), (2) correlated to humidity: sun radiation intensity (correlation of -0.72) and minutes of sunshine every 10 min (correlation of -0.66), (3) not correlated to anything: precipitation, wind speed and wind direction. In addition, it should be mentioned that the sun radiation and minutes of sunshine are also strongly correlated, as both measurements are derivatives of the sun's intensity.

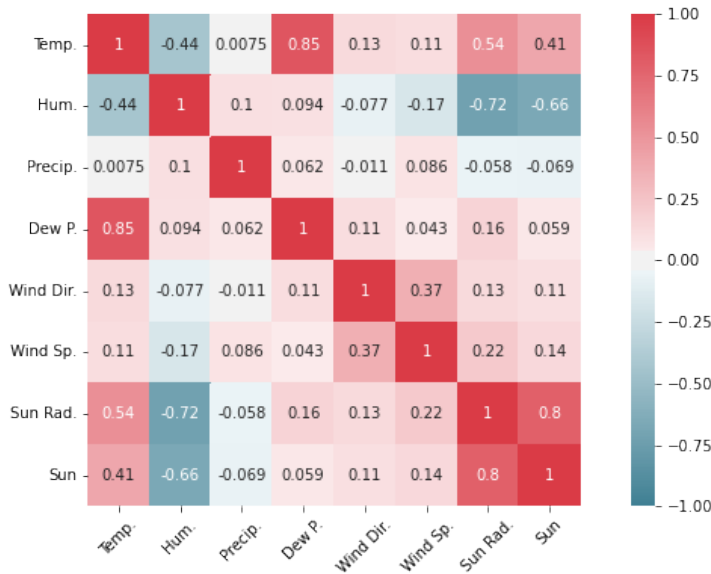


Fig. F.3: Raw weather data correlation matrix. Three categories are formed based on the correlations between the captured conditions—ones significantly correlated to temperature, ones significantly correlated to humidity and ones not significantly connected to anything.

8.2 Data Sampling

Example of a 100-frame sampling created for the training week between 13-20 of February can be seen in Fig. F.4 together with examples of the selected images. The blue points are all the images present in the week of February, while the red ones are the sampled images.

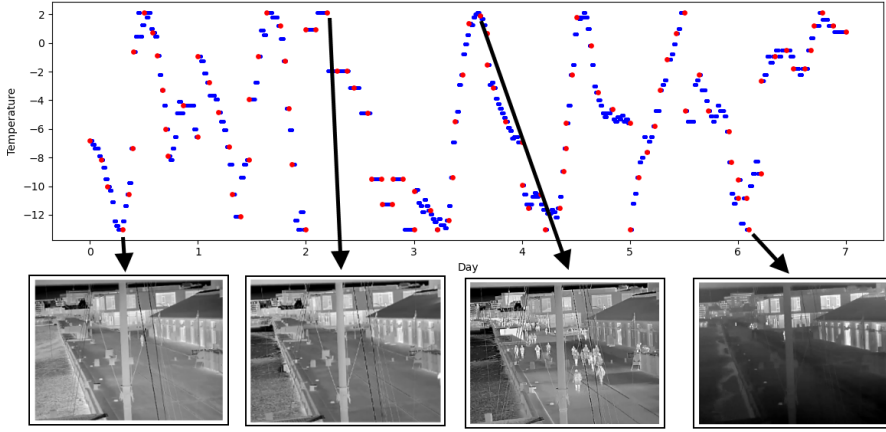
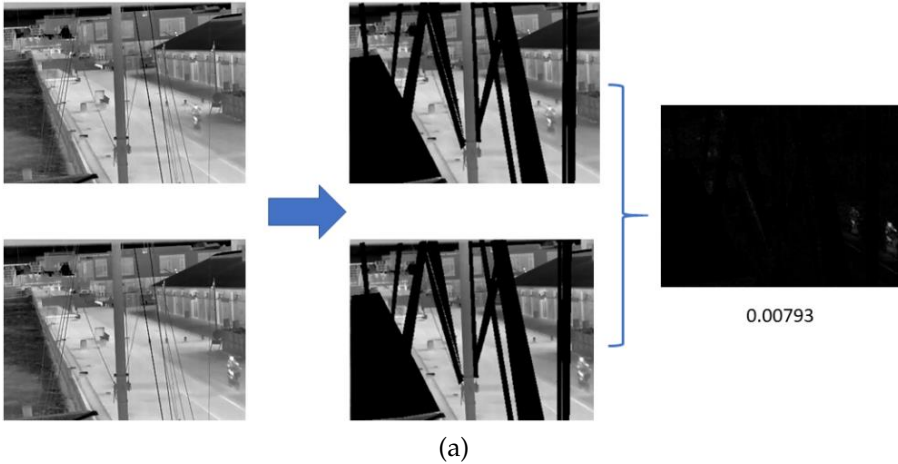


Fig. F.4: Sampling of the week between 13-20 of February. The sampling is done depending on the frame numbering and temperature. Some example images and their positions are given.

8.3 Activity Calculation

The difference between two consecutive frames in the dataset is set as the activity of the scene. When multiple people are moving in the scene or cars and bikes are passing through this will prompt a strong change between consecutive frames, signifying more activity. As the background contains moving parts which can be misinterpreted as activities, a mask is created, removing the waterfront water and the swaying ropes and masts. Examples of the calculation steps for a scene with small activity change and one with large are given in Fig F.5, where the final numbers represent the activity.



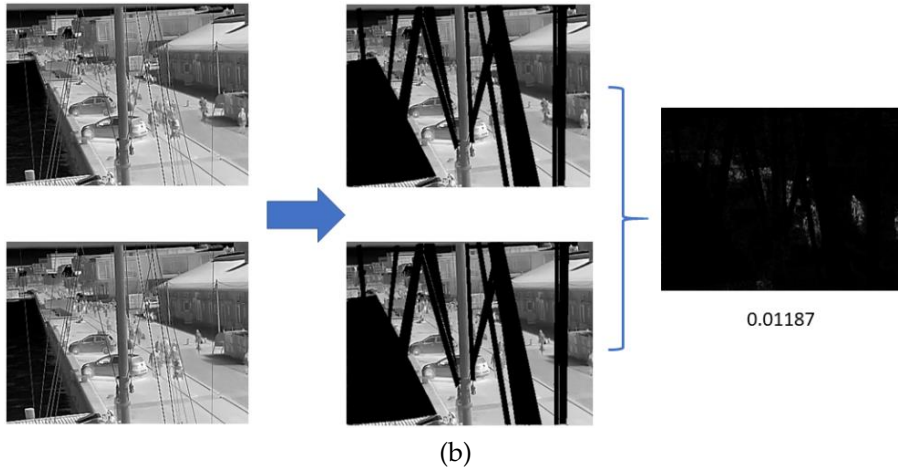


Fig. F.5: Steps for calculating the scene activity level. (a) A frame with low activity. (b) A frame with high activity. The middle images in (a) and (b) represent the masked moving elements of the background—the water, mast and ropes.

References

- [1] Indrė Žliobaitė, Mykola Pechenizkiy, and Joao Gama, “An overview of concept drift applications,” *Big data analysis: new algorithms for a new society*, pp. 91–114, 2016.
- [2] Mingyang Guan, Changyun Wen, Mao Shan, Cheng-Leong Ng, and Ying Zou, “Real-time event-triggered object tracking in the presence of model drift and occlusion,” *IEEE Transactions on Industrial Electronics*, vol. 66, no. 3, pp. 2054–2065, 2018.
- [3] Limin Yang, Wenbo Guo, Qingying Hao, Arridhana Ciptadi, Ali Ahmadzadeh, Xinyu Xing, and Gang Wang, “CADE: Detecting and explaining concept drift samples for security applications,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2327–2344.
- [4] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.
- [5] Waqas Sultani, Chen Chen, and Mubarak Shah, “Real-world anomaly detection in surveillance videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

- [6] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.
- [7] W. Liu, D. Lian W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [8] Meng Wang, Wei Li, and Xiaogang Wang, "Transferring a generic pedestrian detector towards specific scenes," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3274–3281.
- [9] Manzoor Ahmed Hashmani, Syed Muslim Jameel, Hitham Al-Hussain, Mobashar Rehman, and Arif Budiman, "Accuracy performance degradation in image classification models due to concept drift," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, 2019.
- [10] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *arXiv preprint arXiv:1906.02530*, 2019.
- [11] Abhijit Suprem, Joy Arulraj, Calton Pu, and Joao Ferreira, "Odin: automated drift detection and recovery in video analytics," *arXiv preprint arXiv:2009.05440*, 2020.
- [12] Zhiheng Yang, Jun Li, and Huiyun Li, "Real-time pedestrian and vehicle detection for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 179–184.
- [13] Zhilu Chen and Xinming Huang, "Pedestrian detection for autonomous vehicle using multi-spectral cameras," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 211–219, 2019.
- [14] Li Chen, Nan Ma, Patrick Wang, Jiahong Li, Pengfei Wang, Guilin Pang, and Xiaojun Shi, "Survey of pedestrian action recognition techniques for autonomous driving," *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 458–470, 2020.
- [15] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.
- [16] Yahia Fahem Said and Mohammad Barr, "Pedestrian detection for advanced driver assistance systems using deep learning algorithms," *IJC-SNS*, vol. 19, no. 10, 2019.

- [17] Guojiang Shen, Linfeng Zhu, Jihan Lou, Si Shen, Zhi Liu, and Longfeng Tang, "Infrared multi-pedestrian tracking in vertical view via siamese convolution network," *IEEE Access*, vol. 7, pp. 42718–42725, 2019.
- [18] Özgür Göçer, Kenan Göçer, Barış Özcan, Mujesira Bakovic, and M Furkan Kırac, "Pedestrian tracking in outdoor spaces of a suburban university campus for the investigation of occupancy patterns," *Sustainable cities and society*, vol. 45, pp. 131–142, 2019.
- [19] Aske Rasch Lejbølle, Kamal Nasrollahi, Benjamin Krogh, and Thomas B Moeslund, "Person re-identification using spatial and layer-wise attention," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1216–1231, 2019.
- [20] Hui Li, Meng Yang, Zhihui Lai, Weishi Zheng, and Zitong Yu, "Pedestrian re-identification based on tree branch network with local and global learning," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 694–699.
- [21] Hua Han, MengChu Zhou, Xiwu Shang, Wei Cao, and Abdullah Abusorrah, "Kiss+ for rapid and accurate pedestrian re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 394–403, 2020.
- [22] Santhosh Kelathodi Kumaran, Debi Prosad Dogra, and Partha Pratim Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *arXiv preprint arXiv:1901.08292*, 2019.
- [23] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019.
- [24] Fouzi Harrou, Nabil Zerrouki, Ying Sun, and Amrane Houacine, "An integrated vision-based approach for efficient human fall detection in a home environment," *IEEE Access*, vol. 7, pp. 114966–114974, 2019.
- [25] Faten A Elshwemy, Reda Elbasiony, and Mohamed Talaat Saidahmed, "A new approach for thermal vision based fall detection using residual autoencoder," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 2, pp. 250–258, 2020.
- [26] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Dimitris Dres, and Dimitris Drakoulis, "Man overboard event detection from rgb and thermal imagery: Possibilities and limitations," in *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, 2020, pp. 1–6.

- [27] Jinsong Liu, Mark Philip Philipsen, and Thomas B Moeslund, "Supervised versus self-supervised assistant for surveillance of harbor fronts," in *16th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2021.
- [28] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros, "Unsupervised domain adaptation through self-supervision," *arXiv preprint arXiv:1909.11825*, 2019.
- [29] Pravin Nagar, Mansi Khemka, and Chetan Arora, "Concept drift detection for multivariate data streams and temporal segmentation of daylong egocentric videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1065–1074.
- [30] Carlos Mera, Mauricio Orozco-Alzate, and John Branch, "Incremental learning of concept drift in multiple instance learning for industrial visual inspection," *Computers in Industry*, vol. 109, pp. 153–164, 2019.
- [31] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *arXiv preprint arXiv:1810.11953*, 2018.
- [32] Tegjyot Singh Sethi and Mehmed Kantardzic, "Handling adversarial concept drift in streaming data," *Expert systems with applications*, vol. 97, pp. 18–40, 2018.
- [33] Paulo RL Almeida, Luiz S Oliveira, Alceu S Britto Jr, and Robert Sabourin, "Adapting dynamic classifier selection for concept drift," *Expert Systems with Applications*, vol. 104, pp. 67–85, 2018.
- [34] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.
- [35] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James McBride, "Ford multi-av seasonal dataset," *The International Journal of Robotics Research*, vol. 39, no. 12, pp. 1367–1376, 2020.
- [36] Bharathkumar Ramachandra and Michael Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2569–2578.

- [37] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs, "Meva: A large-scale multiview, multimodal video dataset for activity detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1060–1068.
- [38] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López, "Pedestrian detection at day/night time with visible and fir cameras: A comparison," *Sensors*, vol. 16, no. 6, pp. 820, 2016.
- [39] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [40] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al., "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8601–8610.
- [41] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide, "Gated2depth: Real-time dense lidar from gated images," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7374–7383.
- [43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Acdd: The adverse conditions dataset with correspondences for semantic driving scene understanding," *arXiv preprint arXiv:2104.13395*, 2021.
- [44] Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.
- [45] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.
- [46] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al., "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011. IEEE*, 2011, pp. 3153–3160.

- [47] Mantini Pranav, Li Zhenggang, et al., "A day on campus-an anomaly detection dataset for events in a single camera," in *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [48] Ilker Bozcan and Erdal Kayacan, "Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8504–8510.
- [49] FLIR, "Flir thermal dataset for algorithm training," <https://www.flir.com/oem/adas/adas-dataset-form/>, 2019, last accessed: September, 2021.
- [50] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng, "Ptb-tir: A thermal infrared pedestrian tracking benchmark," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 666–675, 2019.
- [51] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2626–2634.
- [52] Yupeng Zhang, Yuheng Lu, Hajime Nagahara, and Rin-ichiro Taniguchi, "Anonymous camera for privacy protection," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 4170–4175.
- [53] Chao Ma, Ngo Thanh Trung, Hideaki Uchiyama, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi, "Adapting local features for face detection in thermal image," *Sensors*, vol. 17, no. 12, pp. 2741, 2017.
- [54] My Kieu, Andrew D Bagdanov, and Marco Bertini, "Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1, pp. 1–19, 2021.
- [55] Yole Développement, "Thermal imagers and detectors 2020 - covid-19 outbreak impact – preliminary report," http://www.yole.fr/Thermal_Imagers_And_Detectors_Covid19_Outbreak_Impact.aspx, 2020, last accessed: August, 2021.
- [56] Allied Market Research, "Global thermal imaging camera market by 2030," <https://www.globenewswire.com/news-release/2021/08/09/2277188/0/en/Global-Thermal-Imaging-Camera-Market-is-Expected-to-Reach-7-49-Billion-by-2030-Says-AMR.html>, 2021, last accessed: August, 2021.

- [57] Mordor Intelligence, "Ir camera market - growth, trends, covid-19 impact, and forecasts (2021 - 2026)," <https://www.mordorintelligence.com/industry-reports/ir-camera-market>, 2021, last accessed: August, 2021.
- [58] Hikvision, "Ds-2td2235d-25/50," <https://us.hikvision.com/en/products/more-products/discontinued-products/thermal-camera/thermal-network-bullet-camera-ds>, 2015, last accessed: May, 2021.
- [59] Danish Meteorological Institute, "Dmi api," <https://confluence.govcloud.dk/display/FDAPI>, 2019, last accessed: May, 2021.
- [60] Quang Huy Tran, Dongyeob Han, Choonghyun Kang, Achintya Haldar, and Jungwon Huh, "Effects of ambient temperature and relative humidity on subsurface defect detection in concrete structures by active thermal imaging," *Sensors*, vol. 17, no. 8, pp. 1718, 2017.
- [61] CA James, AJ Richardson, PW Watt, and NS Maxwell, "Reliability and validity of skin temperature measurement by telemetry thermistors and a thermal camera during exercise in the heat," *Journal of thermal biology*, vol. 45, pp. 141–149, 2014.
- [62] M Ball and Harry Pinkerton, "Factors affecting the accuracy of thermal imaging cameras in volcanology," *Journal of Geophysical Research: Solid Earth*, vol. 111, no. B11, 2006.
- [63] AA Gowen, BK Tiwari, PJ Cullen, K McDonnell, and CP O'Donnell, "Applications of thermal imaging in food quality and safety assessment," *Trends in food science & technology*, vol. 21, no. 4, pp. 190–200, 2010.
- [64] Erwan Bernard, Nicolas Rivière, Mathieu Renaudat, Michel Péalat, and Emmanuel Zenou, "Active and thermal imaging performance under bad weather conditions," 2014.
- [65] JOSEFINE CORNÉ and ULRIKA HELANDER SJÖBLOM, *Investigation of IR transmittance in different weather conditions and simulation of passive IR imaging for flight scenarios*, Ph.D. thesis, MS thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2019.
- [66] DM Bulanon, TF Burks, and V Alchanatis, "Study on temporal variation in citrus canopy using thermal imaging for citrus fruit detection," *Biosystems Engineering*, vol. 101, no. 2, pp. 161–171, 2008.
- [67] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in Neural Information*

- Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.
- [68] Alex McKinney, "Vqvae2 implementation," <https://github.com/vvwm23/vqvae-2>, 2021, last accessed: June, 2021.
 - [69] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.
 - [70] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.
 - [71] Ultralytics, "Yolov5," <https://github.com/ultralytics/yolov5>, 2020, last accessed: April, 2021.
 - [72] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar, "Thermal object detection in difficult weather conditions using yolo," *IEEE Access*, vol. 8, pp. 125459–125476, 2020.
 - [73] Noor Ul Huda, Bolette D Hansen, Rikke Gade, and Thomas B Moeslund, "The effect of a diverse dataset for transfer learning in thermal person detection," *Sensors*, vol. 20, no. 7, pp. 1982, 2020.
 - [74] Yung-Yao Chen, Sin-Ye Jhong, Guan-Yi Li, and Ping-Han Chen, "Thermal-based pedestrian detection using faster r-cnn and region decomposition branch," in *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, 2019, pp. 1–2.
 - [75] Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.
 - [76] Tzutalin, "Labelimg," <https://github.com/tzutalin/labelImg>, 2015, last accessed: June, 2021.
 - [77] Krzysztof Stopa, Andrey Kobyshev, Matthias, and Hadrien Bertrand, "Suntime," <https://github.com/SatAgro/suntime>, 2019, last accessed: July, 2021.
 - [78] Jean Meeus, "Astronomical algorithms," *Richmond*, 1991.

- [79] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al., “Measuring and testing dependence by correlation of distances,” *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.
- [80] Wladston Filho, “Distance correlation,” <https://gist.github.com/wladston>, 2020, last accessed: July, 2021.
- [81] Clive WJ Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.
- [82] Skipper Seabold and Josef Perktold, “statsmodels: Econometric and statistical modeling with python,” in *9th Python in Science Conference*, 2010.
- [83] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox, “Neural granger causality,” *arXiv preprint arXiv:1802.05842*, 2018.
- [84] Neal Parikh and Stephen Boyd, “Proximal algorithms,” *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 127–239, 2014.
- [85] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, “Isolation-based anomaly detection,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 1–39, 2012.
- [86] John Platt et al., “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
- [87] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [88] Jonas MB Haslbeck, Laura F Bringmann, and Lourens J Waldorp, “A tutorial on estimating time-varying vector autoregressive models,” *Multivariate Behavioral Research*, vol. 56, no. 1, pp. 120–149, 2021.
- [89] Vincent Le Guen and Nicolas Thome, “Probabilistic time series forecasting with structured shape and temporal diversity,” *arXiv preprint arXiv:2010.07349*, 2020.
- [90] Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, WEI Ying, and Junzhou Huang, “Adversarial sparse transformer for time series forecasting,” 2020.

References

Paper G

Detecting Anomalies Reliably in Long-term Surveillance Systems

Jinsong Liu, Ivan Nikolov, Mark P. Philipsen, and Thomas B.
Moeslund

The paper has been published in the
*Proceedings of the 17th International Joint Conference on Computer Vision,
Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 2022*

© 2022 SCITEPRESS

DOI: 10.5220/0010907000003124

The layout has been revised.

Abstract

In surveillance systems, detecting anomalous events like emergencies or potentially dangerous incidents by manual labor is an expensive task. To improve this, anomaly detection automatically by computer vision relying on the reconstruction error of an autoencoder (AE) is extensively studied. However, these detection methods are often studied in benchmark datasets with relatively short time duration—a few minutes or hours. This is different from long-term applications where time-induced environmental changes impose an additional influence on the reconstruction error. To reduce this effect, we propose a weighted reconstruction error for anomaly detection in long-term conditions, which separates the foreground from the background and gives them different weights in calculating the error, so that extra attention is paid on human-related regions. Compared with the conventional reconstruction error where each pixel contributes the same, the proposed method increases the anomaly detection rate by more than twice with three kinds of AEs (a variational AE, a memory-guided AE, and a classical AE) running on long-term (three months) thermal datasets, proving the effectiveness of the method.

keywords: surveillance; anomaly detection; autoencoder; long-term; weighted reconstruction error; background estimation

1 Introduction

For a safer daily life, round-the-clock surveillance systems have been installed in some private and public places. Generally they are manually operated, which is expensive. Therefore, an automatic tool to help find emergencies or potentially dangerous incidents that require extra attention is in dire needed.

From the perspective of computer vision, such a tool can be realized using either supervised or unsupervised learning. Supervised learning needs a large amount of annotated data illustrating what the emergencies or potentially dangerous incidents look like. This is too expensive as collecting enough data of rarely-occurring incidents is time consuming and even unfeasible. On the contrary, unsupervised learning greatly lowers the cost, making it more preferred in this task.

This unsupervised solution is often realized by anomaly detection via an autoencoder (AE), which treats these rarely-happening emergencies and potentially dangerous incidents as anomalies but frequently-occurring incidents as normal. In general, an anomaly is deviating from a normal in many aspects. An AE trained with only normal data can reconstruct similar normal patterns with minimal errors, but struggles with abnormal patterns. Hence the difference between the input and the reconstructed output, usually in the form of mean square error (MSE), has the ability to measure the input's de-

viation from the normal data. Input with the MSE larger than a predefined threshold is detected as an anomaly.

This detection strategy works with an assumption that the concept of what is normal is constant. Benchmark datasets on which existing anomaly detection solutions are evaluated satisfy this assumption, because they are relatively short in time duration. However, in real life a surveillance system will be running for months and hence the normal pattern will inevitably drift. This can be illustrated in Fig. G.1 where all these four harbor front scenes are normal in terms of human activities, but the obvious changes across time in contrast, illumination, water ripples, and other environmental aspects make them different from what has been defined as normal in the training phase. This time-induced drift has a large influence on the reconstruction error and thus the anomaly detection is not reliable any more. This phenomenon raises an open research question—how to detect anomalies reliably in long-term surveillance systems.

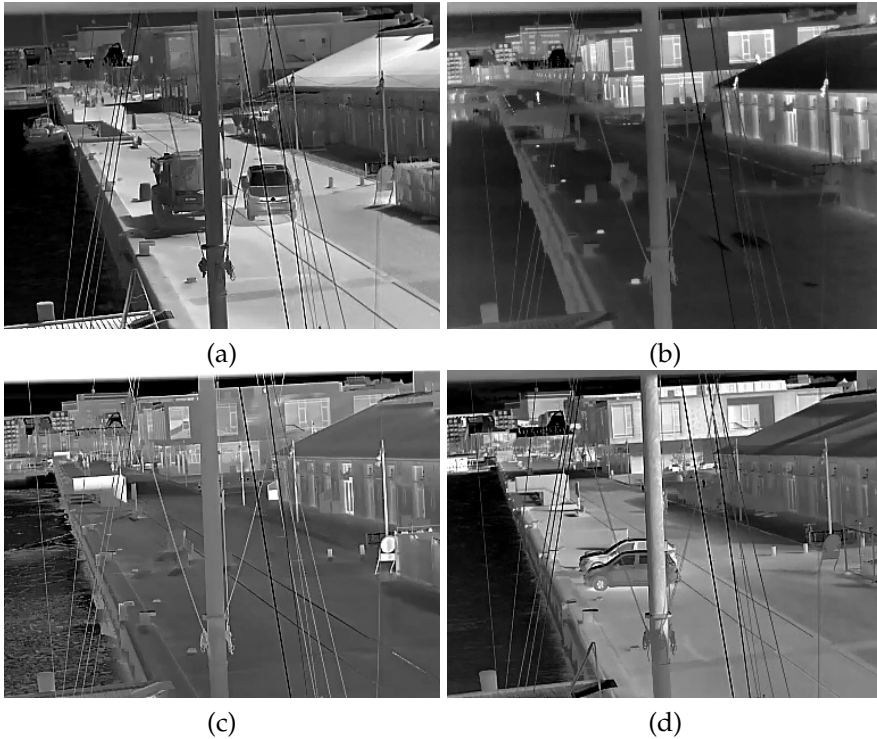


Fig. G.1: Example images from different months. All show normal activity but with significant differences due to the seasons. (a) August. (b) January. (c) February. (d) April.

To this end, we propose a weighted reconstruction error method that uses different weights for foreground pixels and background pixels in calculat-

ing the error, for which five background estimation (foreground extraction) methods are implemented and evaluated. In this way, the influence of the time-induced drift on the reconstruction error is reduced and hence anomaly detection is more reliable.

By applying the proposed method to long-term datasets spanning three months (August 2020, January 2021, April 2021) collected from a real-world harbor front surveillance system, the experimental results show that the weighted reconstruction error increases the anomaly detection rate by at least twice than that with the conventional reconstruction error, for all the three kinds of AEs (a variational AE, a memory-guided AE, and a classical AE), proving the effectiveness of the method.

The datasets and code are published on GitHub—<https://github.com/JinsongCV/Weighted-MSE>, making the integration of the weighted reconstruction error and the comparison between before and after results much easier.

2 Related Work

Existing work on anomaly detection [1–10] is usually AE-based. Though some attempts are made to improve the anomaly detection performance, for example incorporating temporal information [3–5], introducing a generative adversarial network (GAN) to differentiate reconstructions from inputs [6], using both the memorized features of the training set and the input’s features to do reconstruction [7, 11], and so on, these methods are only studies on benchmark datasets—Avenue [12], ShanghaiTech [13], UCSD [14], UMN [15], and Subway [16]), which have an imperfection in common—a short duration of a few minutes or hours (shown in Table G.1) [17, 18]. Therefore, generalizing the existing work evaluated on such datasets to a long-term application in real life can be problematic, considering the extra time-induced changes. For example, the illumination and contrast vary from the shifts in day and night, weather, seasons, etc. This environmental drift imposes an additional variation on the reconstruction error and thus makes it not solely correlated to human activities that are responsible for most anomalies.

Table G.1: Time duration (hours) of benchmark datasets for anomaly detection.

Avenue	ShanghaiTech	UCSD	UMN	Subway
0.5	3.6	3.1	0.07	2.3

This challenge inspires us to focus more on foreground regions where anomalies are assumed in when calculating the reconstruction error, to eliminate the influence of the time-induced environmental drift, which is exactly the proposed weighted reconstruction error does.

A similar solution to ours is the object-centric AE [19, 20] that takes the pre-detected object region instead of the full image as the input. Despite the similarity, there are four distinctions. (i) The goals are different. Their work expects to generalize an AE trained on one scene to another scene without further finetuning, while our method targets to reduce the effect of the environmental drift in long-term surveillance systems. (ii) Our method still reconstructs the full image instead of only object regions, because the location of an object relative to the background is important, for example, a drowning accident only happens in the water area. (iii) Our method is much more flexible like a post-processing module and thus easily incorporated to any framework. (iv) Our method treating foreground and background regions separately also provides an ability to investigate environmental anomalies like a sudden contrast change due to an extreme weather event.

3 Methods

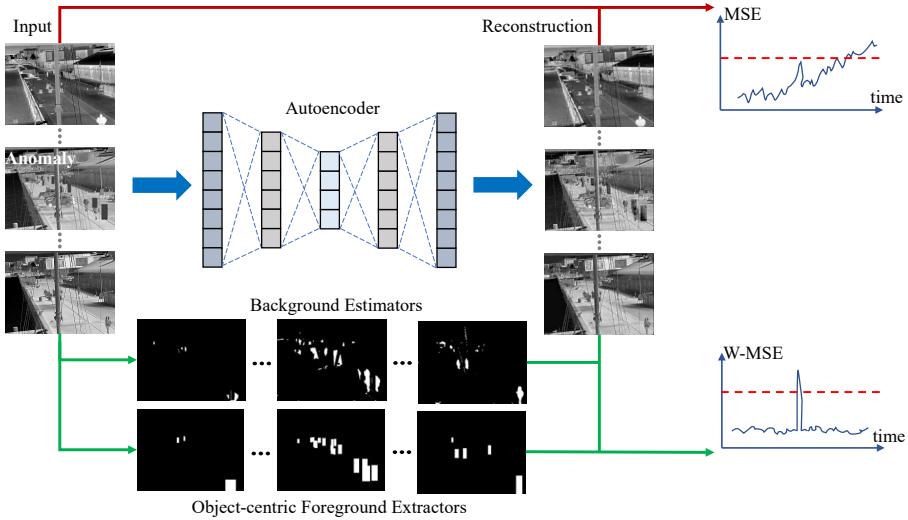


Fig. G.2: Diagram of the proposed method.

This paper proposes a weighted reconstruction error for anomaly detection illustrated by the diagram in Fig. G.2.

In it, the red flow indicates the conventional anomaly detection scheme where the reconstruction error (in the form of MSE) is directly calculated from the input and the reconstructed output with each pixel contributes the same. This calculation also considers the time-induced environmental drift as part of the reconstruction error, and thus for input spanning a long time

period the MSE curve will fluctuate greatly. This is very dangerous as a real anomaly will be ignored, if its MSE value is lower than other fluctuated MSE values of normal inputs. Such a phenomenon is shown in the upper MSE curve where normal inputs with drift have larger MSE values than the threshold (the red dashed line), not only introducing false positives but also missing the real anomaly.

In contrast, the green flow in the diagram indicates the proposed weighted reconstruction error-based anomaly detection scheme. Additional background estimators or object-centric foreground extractors can segment an input into foreground region and background region. This information together with the input and the reconstruction are used to calculate the reconstruction error where the foreground pixels and background pixels are assigned different weights, so that the error focuses more on the region where anomalies usually happen and thus the effect of the environmental drift is reduced. In this way, the weighted MSE curve will be much more smoother for normal inputs but generates a peak if an anomaly comes in, like the lower W-MSE curve shows. Another thing to be mentioned is that both the red flow scheme and the proposed green flow scheme are indicating the inference phase—*anomaly detection*.

3.1 Autoencoder

Following what is customary, we use an AE to detect anomalies by finding frames with the largest reconstruction errors. Three AEs are applied. The first is a variational AE—VQVAE2 [21] whose encoder compresses the input into multi-scale quantized latent maps for the decoder to process. The second is a memory-guided AE—MNAD [11] that uses a concatenated latent space (of the naive latent space from the encoder output and the typical features stored in a memory module constructed from training) to reconstruct the input. An anomaly is measured by not only the reconstruction error but also the distance between the encoder output and the nearest memorized features. The third is a classical AE (CAE) designed by us, which is without any advanced processing of the latent space. This CAE uses eleven convolution layers and five pooling layers to downsize the input ($384 \times 288 \times 1$) into a compressed feature tensor ($10 \times 7 \times 64$), and another six transposed convolution layers and five convolution layers to transform the latent feature space into the reconstructed output. Detailed implementations are shared on GitHub.

3.2 Background Estimation

As mentioned before, the method is characteristic of foreground regions and background regions contributing differently to the weighted reconstruction

error. Therefore, separating the background from the foreground is the key. To achieve this we test out two pipelines, one using classical statistical methods to estimate the background, the other one using the result of a human detector as the foreground and everything else as the background.

In section 4, we will test all the methods of the two pipelines and determine which is the best method or the best combination of a few methods to separate the foreground from the background, so that the drift can be removed effectively for improving the anomaly detection rate.

Statistical Background Estimation

This pipeline is composed of classical statistical approaches instead of deep learning segmentation methods [22, 23] to minimize the complexity. Also this avoids the high price of supervised segmentation concerning pixel-level annotations. In our harbor front scenario, the objects in the foreground vary significantly—humans from a single one to groups, vehicles, bicycles, and others. These variations cause extra difficulties and manpower in pixel-level annotations if a deep model is chosen. The four classical background estimators are as follows:

- Mixture of Gaussians (MOG2) [24] — using Gaussian mixture probability density to continuously model the background.
- Mixture of Gaussians using K-nearest neighbours (KNN) [25] — an extension of the MOG2 method by implementing a K-nearest neighbours algorithm on top for a more robust kernel density estimation.
- Image difference with arithmetic mean (ID_a) — the difference between the current image and the previous one is processed by adaptive thresholding to get the background mask. ID_a uses an arithmetic mean weight—each pixel in the neighborhood contributes equally to compute the local threshold.
- Image difference with Gaussian mean (ID_g) — the same principle with ID_a , but with a different adaptive thresholding strategy. ID_g uses a Gaussian mean weight—pixels in the neighborhood farther away from the center contribute less to the local threshold computing.

To do background estimation, all the four methods need the neighbouring images of the current frame. For the MOG2 and KNN methods, the number of neighbouring images is heuristically set to 20, as it has been shown that more frames are better at modeling the background. For the ID_a and ID_g methods, only one previous image is used.

Once a mask is acquired from any of the four methods, it goes through a post-processing procedure—a morphological closing with a structuring element of size 7×7 followed by an opening with an element of size 3×3 . This step serves to remove small noise particles. Finally, the moving elements in the background like the water, ropes, and masts are removed from the mask by prior knowledge of their locations. The resulting mask will have foreground pixels with large grayscale values approaching 255 and background pixels with small values near 0. All of these procedures are implemented from the OpenCV library [26].

Object-centric Foreground Extraction

Besides the above four classical approaches, we test another method—object-centric foreground extraction, provided that there is a well-trained human detector at hand and human activities are the targets. The detector we use is YOLOv5 [27, 28], with which each person is represented by a rectangle in the mask. The pixels in the rectangle has a same grayscale value—the person’s detection confidence multiplied by 255, while pixels in other regions are with the value 0.

As a whole, these five versions of masks explicitly locate foreground areas with very large grayscale values, so for a clear reference the subsequent contents will call such a mask foreground map. Fig. G.3 shows one input image and the results from the five methods.

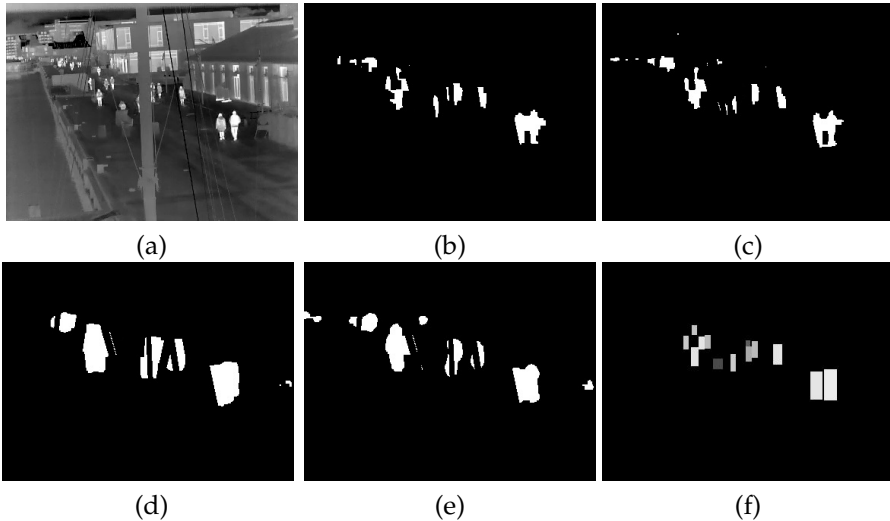


Fig. G.3: An input thermal image (a) and the outputs (b)-(f) from the five implemented background estimation (foreground extraction) methods. (a) Input. (b) MOG2. (c) KNN. (d) ID_a . (e) ID_g . (f) YOLOv5.

3.3 Weighted Reconstruction Error

First to be noted is that this paper goes with the convention and thus opts for the MSE to measure the difference between the input and the reconstructed output, so the following contents will directly use MSE to represent the difference without further explanation.

As long as there is a foreground map M locating foreground pixels P_{fg} , any weighted MSE is possible by giving P_{fg} and background pixels P_{bg} arbitrarily-defined weights for a specific task.

For an input I with size $H \times W$ and its corresponding reconstruction R , a general weighted MSE_w is:

$$MSE_w = \frac{\sum_{i=1}^H \sum_{j=1}^W (I_{ij} \cdot \bar{M}_{ij} - R_{ij} \cdot \bar{M}_{ij})^2}{H \times W} \quad (G.1)$$

where \bar{M}_{ij} is the value of the weight map \bar{M} at pixel (i, j) . \bar{M} is calculated using Equation G.2.

$$\bar{M} = \frac{w_{fg} \times M + w_{bg} \times (255 - M)}{255} \quad (G.2)$$

where w_{fg} and w_{bg} are normalized weights for P_{fg} and P_{bg} , respectively; as an 8-bit image, $255 - M$ is the “inverse” operation of M , explicitly locating P_{bg} ; therefore \bar{M} is the final weight map normalized to 0-1 for calculating weighted MSE in Equation G.1.

Setting w_{fg} as 1 and w_{bg} as 0 is the special case of the MSE only considering P_{fg} . Likewise, setting w_{fg} as 0 and w_{bg} as 1 is the MSE only looking at P_{bg} .

A more general case is a weighted MSE_w combining foreground maps (e.g., M_1, M_2) from several background estimators (foreground extractors).

$$MSE_w = \frac{\sum_{i=1}^H \sum_{j=1}^W (I_{ij} \cdot \bar{M}_{ij} - R_{ij} \cdot \bar{M}_{ij})^2}{H \times W} \quad (G.3)$$

$$\bar{M} = w_1 \times \bar{M}_1 + w_2 \times \bar{M}_2 \quad (G.4)$$

$$\bar{M}_1 = \frac{w_{fg1} \times M_1 + w_{bg1} \times (255 - M_1)}{255} \quad (G.5)$$

$$\bar{M}_2 = \frac{w_{fg2} \times M_2 + w_{bg2} \times (255 - M_2)}{255} \quad (G.6)$$

where, \bar{M}_1 is the weight map from M_1 with w_{fg1} and w_{bg1} as normalized weights for P_{fg} and P_{bg} in M_1 ; \bar{M}_2 is the weight map from M_2 with w_{fg2} and w_{bg2} as normalized weights for P_{fg} and P_{bg} in M_2 ; \bar{M} is the final weight map combining \bar{M}_1 with weight w_1 and \bar{M}_2 with weight w_2 ; the resulting MSE_w is the weighted MSE considering foreground maps from two methods.

4 Experiments

4.1 Dataset Information

Two datasets collected from a long-term harbor front surveillance system are used to investigate the proposed weighted MSE on anomaly detection.

One dataset called 300Ver has 300 images with every 100 sampled from August 2020, January 2021, and April 2021, making itself a dataset spanning 76 days. This dataset is a subset of a larger one covering 8-month and publicly available as part of the publication [18]. The sampling protocol for 300Ver is also given in [18] which uses the temperature as a basis to construct datasets covering cold, warm, and in-between months.

The other dataset called 3515Ver is also a subset of the dataset from [18], and has 3515 images intensively sampled with a frame rate of 0.5fps from 15 pm to 18 pm from 14-16 August 2020, 14-16 January 2021, and 14-16 April 2021. This sampling protocol comes from three strategies. (i) Empirically 15 pm to 18 pm is the time period when there are most people present in view. (ii) Three days from each month not only guarantee the data diversity across time but also limit the amount of the dataset for a better visualization. (iii) 0.5fps limits the amount of 3515Ver, at the same time keeping the information continuity between neighboring frames.

In 300Ver persons are annotated with bounding boxes. Therefore, six foreground maps from MOG2, KNN, ID_a , ID_g , YOLOv5, and ground truth (GT), are prepared for each image. The 3515Ver dataset has no such annotations, so only five kinds of foreground maps are calculated.

First the 300Ver dataset is used. The 3515Ver dataset is then used to verify what has been found on 300Ver and the related contents are in the section of extended experiments. There are three reasons why we do experiments on both datasets. (i) 300Ver covers 76 days with less images while 3515Ver have more images but only covering 9 days; these two datasets compensate for each other, making the experiments consider both a long-term duration and a large number of images. (ii) This separation of two datasets avoids the problem that if all the images are sampled intensively from the 76 days, the resultant 30000 images will make the visualization of drawing the MSE values of them into one curve (like the curve in the following contents) extremely difficult. (iii) Annotating a small dataset (300Ver) is much easier to provide a very accurate foreground extraction, based on which the findings of section 4.3 will be more convincing.

4.2 Implementation Details

Both VQVAE2 and MNAD are trained with 4000 images and validated with 1000 images. CAE is trained with 15000 images and validated with 5000

images due to its naive function compared with the other two. VQVAE2 is trained with a batch size of 32 and a learning rate of 0.0001. MNAD is trained with a batch size of 32, a learning rate of 0.0002, and a value of 0.1 for the weight of the feature separateness and compactness loss. CAE is trained with a batch size of 16 and a learning rate of 0.0003. The training phases stop at the 100th epoch, the 100th epoch, and the 200th epoch for VQVAE2, MNAD, and CAE, respectively, at which the networks are converged with the training losses not decreasing any more. All these training and validation sets are sampled from February 2021 to not only avoid the overlapping with the three-month datasets this paper uses but also enhance the effect of the time-induced drift that we want to address. A kind reminder is that the following experiments are done with all these three AEs but we usually only show related visualizations of VQVAE2 to avoid the repeat of similar results.

The YOLOv5 detector uses a pretrained model from [28] and the training set has no overlapping with the images we use in this paper.

4.3 Weighted MSE

Weighted MSE Curves

To simplify the work and directly answer the question how the conventional MSE and weighted MSE behave for long-term datasets, according to Equation G.1 and Equation G.2, the MSE investigated will consider three situations: the foreground only, the background only, and the full image where each pixel contributes the same as the convention, which are represented as MSE_{fg} , MSE_{bg} , MSE_{cvt} , respectively. These representations will be used in all the following contents.

Therefore, for each AE with 300Ver as input, six kinds of foreground maps produce six MSE_{fg} curves and six MSE_{bg} curves describing the weighted MSE changes across time; likewise, one MSE_{cvt} curve can be drawn to describe the conventional MSE changes across time.

For a better comparison, Fig. G.4 shows the above mentioned 13 MSE curves, produced by the VQVAE2 model. This visualization (of showing multiple curves in one chart) is achieved with a critical pre-processing module before plotting: first the original MSE values are smoothed by a mean filter with its kernel size as 10; then the smoothed values are normalized between 0 and 1; after normalization the curves are overlapped with each other, so a further translation is done for each curve by adding an extra value. In this way, the ranges of curves of MOG2, KNN, ID_a , ID_g , YOLOv5, and GT are [2.5, 3.5], [2.0, 3.0], [1.5, 2.5], [1.0, 2.0], [0.5, 1.5], [0, 1], respectively; the range of the conventional MSE curve is [3.0, 4.0].

From Fig. G.4 several observations are found. (i) The six MSE_{fg} curves in (a) have totally different trends with the trends of MSE_{bg} curves in (b),

4. Experiments

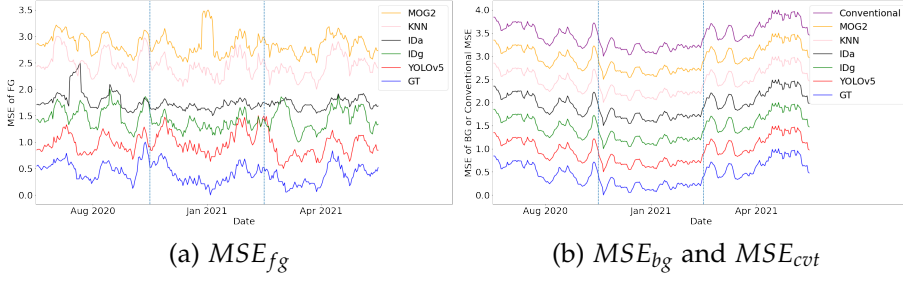


Fig. G.4: MSE (after smoothing, normalization, and translation) curves across time from VQ-VAE2 on 300Ver. The vertical azure dashed lines are used to separate different months.

which is reasonable as the image regions they look at are not the same. (ii) The MSE_{cvt} curve in (b) has almost exactly same trend with that of the six MSE_{bg} curves in (b), but largely deviates from the trends of MSE_{fg} curves in (a), proving that on 300Ver the conventional MSE (where each pixel in the full image contributes the same) cannot represent what happens in the foreground region and thus has no ability to do anomaly detection reliably. (iii) Though the six MSE_{fg} curves in (a) are diverse, they share a similar trend to some extent especially between the MSE_{fg} curve of YOLOv5 and the MSE_{fg} curve of GT. This reflects that they have the ability to represent the foreground changes along with time but also have their own focuses shown by distinct peaks due to the methods' differences. The MSE_{fg} curve of YOLOv5 and the MSE_{fg} curve of GT are bounding box-based focusing only on persons, therefore a larger similarity is found between them. (iv) The trends of all the MSE_{bg} curves and the MSE_{cvt} curve in (b) are U-shape, revealing the influence of the drift across time on the MSE as mentioned before. However, the U-shape trend is not shown in foreground MSE curves in (a), indicating that the time-induced effect influences background regions higher than foreground regions. Hence researches on long-term datasets (applications) need separate analysis on them.

In addition to this, experiments done with MNAD and CAE also get similar results that all support the above findings. As a whole, this part confirms that in long-term datasets (applications) with time-induced drift, the conventional MSE (where each pixel contributes the same) is not suitable to describe the foreground information, not to mention a further step—detecting anomalies.

Weighted MSE for Anomaly Detection

This section will test whether the proposed weighted MSE performs better in anomaly detection. Since there are no specified anomalies in the dataset, and detecting specific anomalies is not the focus of this work, we decide to use

a strategy that maximizes the difference between an anomaly and a normal image, to better focus on the main research problem—how to do anomaly detection reliably in long-term datasets.

To realize this, we synthesize anomalies by overlapping “black-white-pixel” patterns (that the three AEs have never seen) on the person regions of some images. But it seems that such patterns overlapped on only person regions will give the YOLOv5-based foreground map a biased advantage. Hence, to evaluate the five kinds of foreground maps more fairly, four shapes (rectangle, square, circle, and ellipse) of the “black-white-pixel” pattern are considered for the reason that the detector-based map has no round-cornered foregrounds but the other four kinds of maps have. We admit this four-shape strategy cannot totally remove the bias on the YOLOv5-generated map, but if we put the “black-white-pixel” pattern on other foreground regions where there are no people, a greater bias will be given to other statistical background estimators because YOLOv5 only predicts human regions. Therefore, this four-shape strategy should be the best solution to treat these five kinds of foreground maps equally.

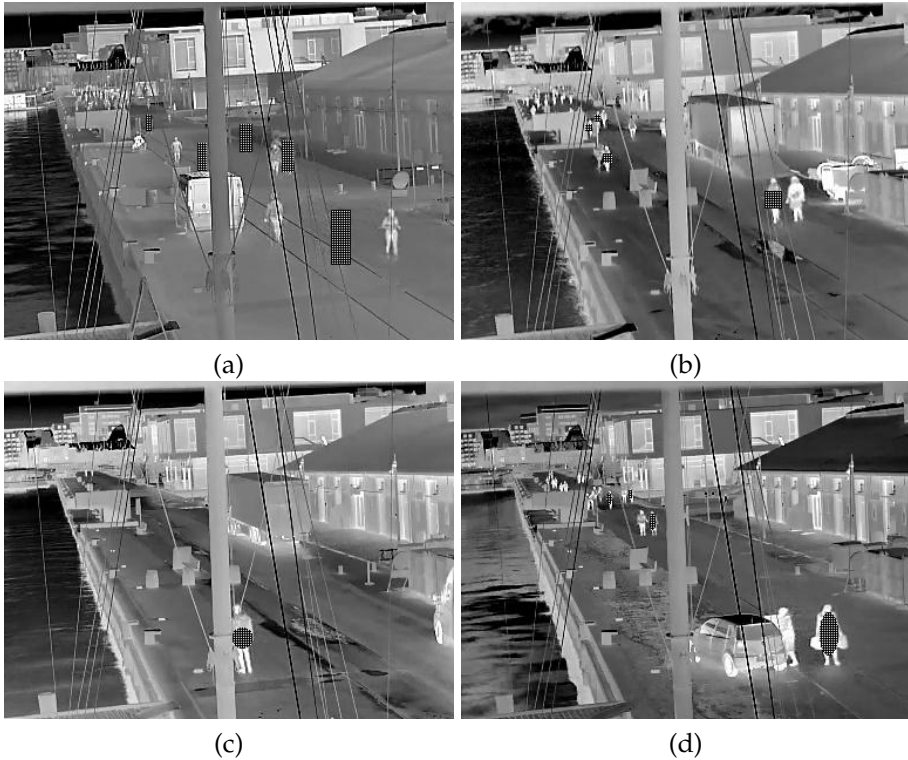


Fig. G.5: Examples of anomalies with “black-white-pixel” patterns in four different shapes. (a) Rectangle. (b) Square. (c) Circle. (d) Ellipse.

4. Experiments

Accordingly, on the premise of having at least one person in each synthesized anomalous image, 21 rectangle-shaped anomalies (the 1st, 11st, 21st, ..., 281st images of 300ver), 21 square-shaped anomalies (the 3rd, 13rd, ..., 293rd images of 300ver), 20 circle-shaped anomalies (the 2nd, 12nd, ..., 282nd images of 300ver), and 16 ellipse-shaped anomalies (the 4th, 14th, ..., 284th images of 300ver) are synthesized. In each of them the annotated GT person regions are randomly chosen to be overlapped with “black-white-pixel” patterns. Examples of the synthesized anomalies are shown in Fig. G.5.

First, the rectangle-shaped anomalies are used to test the anomaly detection rate. Accordingly, in Fig. G.6, six MSE curves (five MSE_{fg} curves and one MSE_{cvt} curve) of VQVAE2 are drawn in color blue, and the anomalies are located with orange peaks. Each sub-figure caption has the same meaning with what has been used in the previous subsection.

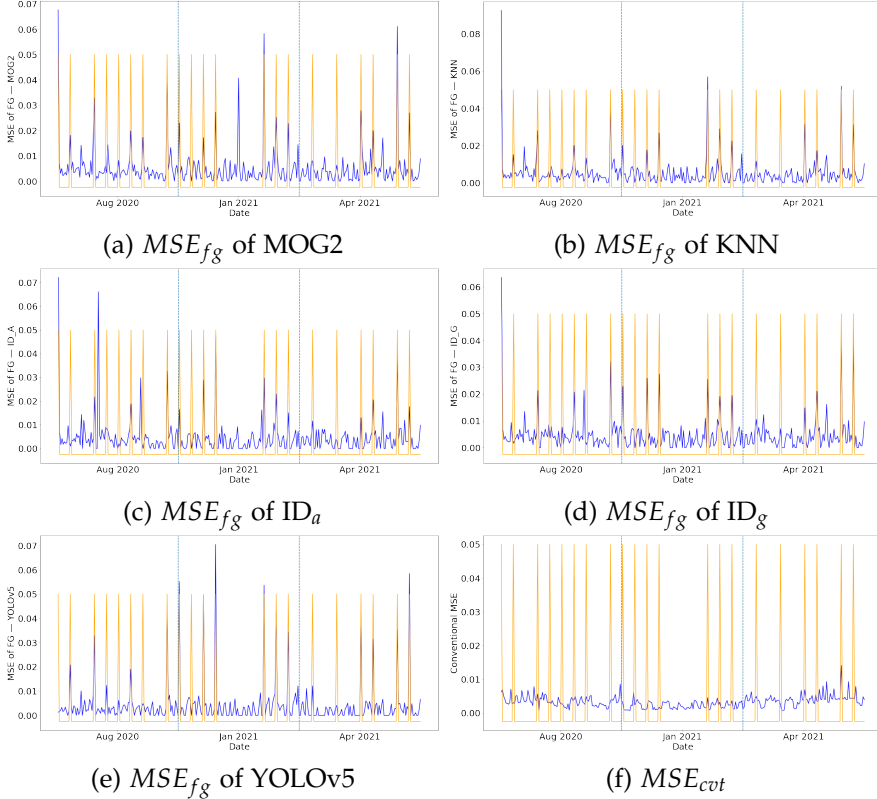


Fig. G.6: After introducing rectangle-shaped anomalies, MSE curves across time from VQVAE2 on the 300Ver dataset. The blue curves describe the MSE changes, and the orange peaks indicate the locations of anomalies. The vertical azure dashed lines are used to separate different months.

From Fig. G.6, the large percentage of overlapping between orange peaks and blue peaks in (a)-(e) proves the usefulness of the proposed weighted MSE in anomaly detection. This also happens in the experiments of VQ-VAE2 on 300Ver but with anomalies in the other three shapes. Specifically, among the images of the largest 30 (10% of the dataset) MSE values of each curve, the number of anomalies is listed in Table G.2. From the table, the weighted MSE using any foreground map has a way high detection rate than the conventional MSE.

Table G.2: Anomaly detection results of weighted MSE and conventional MSE.

	Statistical Background				Object-centric Foreground	Conventional
	MOG2	KNN	ID _a	ID _g	YOLOv5	
Rectangle (21)	16	17	15	15	16	4
Square (21)	13	14	12	13	15	7
Circle (20)	11	13	10	10	16	4
Ellipse (16)	14	12	14	13	14	4
Sum (78)	54	56	51	51	61	19
Detection Rate	69.23%	71.79%	65.38%	65.38%	78.21%	24.36%

When taking multiple foreground maps from different methods into consideration, the top two results in Table G.2—YOLOv5 and KNN, inspire us to combine their foreground maps by applying Equation G.3-G.6 in which w_1 (namely w_{YOLOv5}) and w_2 (namely w_{KNN}) are 0.52 and 0.48, respectively as the normalized values of 78.21% and 71.79%. To be noted is that any combination is possible no matter whether a supervised human detector is available.

To avoid being one-sided, we do further experiments with MNAD and CAE on 300Ver in a way of using rectangle-shaped anomalies and the foreground map combining YOLOv5 and KNN. By using the weighted MSE instead of the conventional MSE, the detection rate increases from 9.52% to 66.67% for MNAD and from 4.76% to 66.67% for CAE.

As a whole, the proposed weighted MSE improves anomaly detection rate markedly on 300Ver—VQVAE2 (2.68 times-3.21 times), MNAD (7 times), CAE (14 times), verifying that this strategy is worth being incorporated in datasets or applications spanning a long time period.

Extended Experiments

The extended experiments on 3515Ver use rectangle-shaped “black-white-pixel” patterns overlapping on the persons who are near the horizontal edge of the water to simulate the anomalies. The resultant 60 synthesized anomalies are consecutive frames and the persons overlapped with the abnormal pattern are fixed individuals. This increases the authenticity of the simulated anomalies—in real life an anomaly usually persists through multiple frames and involves fixed persons.

4. Experiments

Fig. G.7 gives the MSE curves of the three AEs on 3515Ver with synthesized anomalies, in which the curves of absolute MSE values are in blue and the smoothed ones are in red, and the anomalies are located with orange peaks. In Fig. G.7, by using the weighted MSE with the foreground map $M_{YOLOv5\&KNN}$ combining YOLOv5 and KNN, the ascending peaks in (a), (c), and (e) accurately detect the anomalies, yet the conventional MSE curves in (b), (d), and (f) are entirely dominated by time-induced influences for example the fall of a cliff due to the seasonal transition between August 2020 and January 2021. We therefore believe that the extended experiments on a much larger dataset also prove the effectiveness of the proposed weighted MSE in anomaly detection.

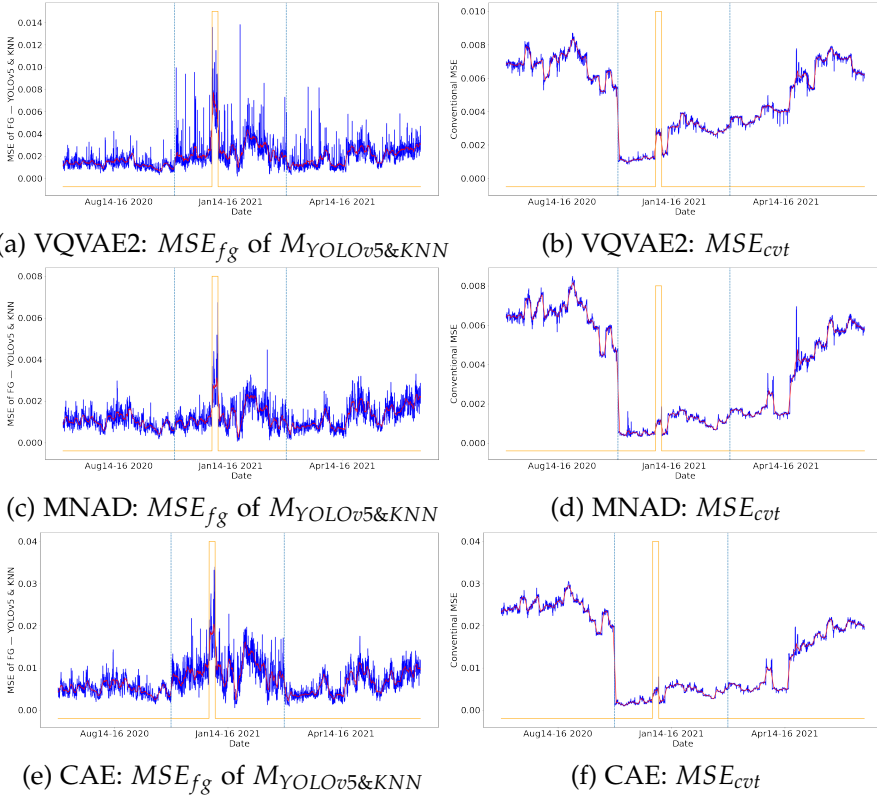


Fig. G.7: MSE curves of VQVAE2, MNAD, and CAE on 3515Ver with synthesized rectangle-shaped anomalies. The curve of absolute MSE values is in blue. The curve of the smoothed values are in red. The anomalies are located with orange peaks. The vertical azure dashed lines are used to separate different months.

5 Conclusions

This paper proposes a weighted reconstruction error in autoencoder-based anomaly detection for long-term surveillance systems. The method aims to make the calculated error more focused on the region where anomalies are assumed in and thus reduces the influence of time-induced environmental drift.

We apply three selected autoencoders to three-month datasets to test the anomaly detection performance. With synthesized anomalies, the autoencoder with proposed weighted reconstruction error always gets a much higher detection rate (more than twice) than the conventional reconstruction error version where each pixel contributes the same, which proves the usefulness of the proposed strategy.

This method is implemented as a flexible module, therefore we expect it can be integrated into and verified by more frameworks. Besides, as a study at harbor fronts, in the future we will use this method to detect emergencies and potentially dangerous incidents like traffic accidents, drowning accidents, crowds in coronavirus days, etc., so that timely controls or rescues by polices, safeguards, and other professionals can be provided for a safer life.

References

- [1] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.
- [2] Yong Shean Chong and Yong Haur Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.
- [3] Jiangpeng Fu, Wentao Fan, and Nizar Bouguila, "A novel approach for anomaly event detection in videos based on autoencoders and se networks," in *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE, 2018, pp. 179–184.
- [4] Haichun Yue, Shigang Wang, Jian Wei, and Yan Zhao, "Abnormal events detection method for surveillance video using an improved autoencoder with multi-modal input," in *Optoelectronic Imaging and Multimedia Technology VI*. International Society for Optics and Photonics, 2019, vol. 11187, p. 111870U.

- [5] Trong-Nguyen Nguyen and Jean Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1273–1283.
- [6] Hao Song, Che Sun, Xinxiao Wu, Mei Chen, and Yunde Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Transactions on Multimedia*, 2019.
- [7] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.
- [8] K Deepak, S Chandrakala, and C Krishna Mohan, "Residual spatiotemporal autoencoder for unsupervised video anomaly detection," *Signal, Image and Video Processing*, pp. 1–8, 2020.
- [9] Du-Ming Tsai and Po-Hao Jen, "Autoencoder-based anomaly detection for surface defect inspection," *Advanced Engineering Informatics*, vol. 48, pp. 101272, 2021.
- [10] Jie Liu, Kechen Song, Mingzheng Feng, Yunhui Yan, Zhibiao Tu, and Liu Zhu, "Semi-supervised anomaly detection with dual prototypes autoencoder for industrial surface inspection," *Optics and Lasers in Engineering*, vol. 136, pp. 106324, 2021.
- [11] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.
- [12] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.
- [13] Weixin Luo, Wen Liu, and Shenghua Gao, , in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.
- [14] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1975–1981.
- [15] Ramin Mehran, Alexis Oyama, and Mubarak Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.

- [16] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, 2008.
- [17] Mantini Pranav, Li Zhenggang, et al., "A day on campus-an anomaly detection dataset for events in a single camera," in *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [18] Ivan Adriyanov Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Velling Dueholm, Anders Skaarup Johansen, Kamal Nasrollahi, and Thomas B Moeslund, "Seasons in drift: A long-term thermal imaging dataset for studying concept drift," in *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.
- [19] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7842–7851.
- [20] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah, "A scene-agnostic framework with adversarial training for abnormal event detection in video," *arXiv e-prints*, pp. arXiv–2008, 2020.
- [21] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in neural information processing systems*, 2019, pp. 14866–14876.
- [22] Mohammadreza Babaei, Duc Tung Dinh, and Gerhard Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.
- [23] Thangarajah Akilan, Qingming Jonathan Wu, Amin Safaei, Jie Huo, and Yimin Yang, "A 3d cnn-lstm-based image-to-image foreground segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 959–971, 2019.
- [24] Zoran Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004. ICPR 2004. IEEE, 2004, vol. 2, pp. 28–31.
- [25] Zoran Zivkovic and Ferdinand Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, no. 7, pp. 773–780, 2006.

References

- [26] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [27] Ultralytics, "Yolov5," <https://github.com/ultralytics/yolov5>, 2020, last accessed: October, 2021.
- [28] Jinsong Liu, Mark P Philipsen, and Thomas B Moeslund, "Supervised versus self-supervised assistant for surveillance of harbor fronts.," in *VISIGRAPP (5: VISAPP)*, 2021, pp. 610–617.

References

Paper H

Imitating Emergencies: Generating Thermal Surveillance Fall Data Using Low-Cost Human-like Dolls

Ivan Nikolov, Jinsong Liu, and Thomas B. Moeslund

The paper has been published in the journal of
Sensors

© 2022 Authors

The layout has been revised.

Abstract

Outdoor fall detection, in the context of accidents, such as falling from heights or in water, is a research area that has not received as much attention as other automated surveillance areas. Gathering sufficient data for developing deep-learning models for such applications has also proven to be not a straight-forward task. Normally, footage of volunteer people falling is used for providing data, but that can be a complicated and dangerous process. In this paper, we propose an application for gathering thermal images of a low-cost rubber doll falling in a harbor, for simulating real emergencies. We achieve thermal signatures similar to a human on different parts of the doll's body. The change of these thermal signatures over time is measured, and its stability is verified. We demonstrate that, even with the size and weight differences of the doll, the produced videos of falling have a similar motion and appearance to what is expected from real people. We show that the captured thermal doll data can be used for the real-world application of pedestrian detection by running the captured data through a state-of-the-art object detector trained on real people. An average confidence score of 0.730 is achieved, compared to a confidence score of 0.761 when using footage of real people falling. The captured fall sequences using the doll can be used as a substitute to sequences of people.

keywords: thermal cameras; fall detection; thermal mannequin; anomaly detection; machine learning

1 Introduction

Automated security systems are becoming increasingly ubiquitous together with the growing requirements for public safety. The possibility to offload parts of the manual surveillance from people to an automated system has driven a great deal of research in better algorithms and extended them for different use cases. These use cases can be roughly separated into outdoor and indoor use. Indoor surveillance mostly focuses on the care for children, patients and elderly [1, 2].

Outdoor surveillance is directed towards detecting suspicious and anomalous pedestrian behaviors on streets, in airports, in libraries [3, 4] and in traffic surveillance and for the early prevention of accidents [5, 6]. Thermal images are also becoming more prevalent for surveillance use cases [7], and this trend is likely to continue to rise in future years due to privacy concerns [8].

Fall detection, as apart of surveillance, has become an increasingly researched field, focusing on methods for the detection and prevention of fall accidents [9]. Fall events can lead to serious injuries and negative consequences to vulnerable groups, such as small children, patients in hospital care and the elderly. It has been shown that a quick reaction is required

when such events occur and especially for the elderly, which can prevent more severe outcomes [10]. Modern deep-learning methods have been employed steadily in tackling the problem of fall detection [11–13].

Such methods have been proven useful in environments with single people as well as crowded areas with a large amount of foot traffic and groups of people [14, 15]. Such deep-learning solutions require enough training and testing data to correctly detect falls and minimize the possibility of either missing or misclassifying important events. Most of the time, the required fall events can occur sporadically or would require certain conditions. This requires data gathering for these events that either depends on acting these scenes out in real life with volunteers [16, 17] or on creating simulated videos and images [18, 19].

Both of these have their pros and cons; however, generally computer simulations cost less time and resources and can produce higher amount of variations but are limited to the scenarios that can be convincingly simulated. For scenarios that cannot be easily simulated, acting them out in real life is left as the only possibility. A problem that can arise in these cases is if the scenario is too dangerous or too complicated for repeated testing. In these cases, mannequins or crash test dummies can be used instead [20, 21]. Ensuring that the used mannequin properly represents a human is then required. This means that the shape, size, appearance, weight, movement and interactions with the environment of the dummy need to be as close to that of humans as possible.

This paper focuses on falls happening in outdoor environments. Detecting falls into water is required as part of drowning prevention surveillance applications. Every year an estimated of 236,000 people drown around the world [22]. Compared to other types of fall events, falls in water are not that widely studied; and datasets and the methods for capturing data are not thoroughly documented. Most of the research is also directed towards specific more-isolated use cases, such as drowning prevention in swimming pools [23, 24] or falls from boats [11], and only some look into warning systems for every day surveillance [25, 26].

As drowning events can happen any time around the clock and systems need to be able to register them independent of weather or lighting conditions, we choose to focus on thermal data. Thermal images are more robust to environmental changes compared to RGB images and can also provide better anonymity preservation, as facial features and clothing cannot be easily recognized. We investigate the use of low-cost dolls for gathering training and testing data for fall detection using thermal images. We look into the requirements for designing a simple doll and separate these requirements into two main categories—appearance and motion.

As we focus on generating thermal data, the appearance requirements include not only the shape of the doll but also the need for it to provide a

temperature distribution similar to the one from real people. The motion requirements revolve around ensuring that constructing falls in a believable way, similar to a human. An overview of the investigation presented in the paper is given in Fig. H.1.

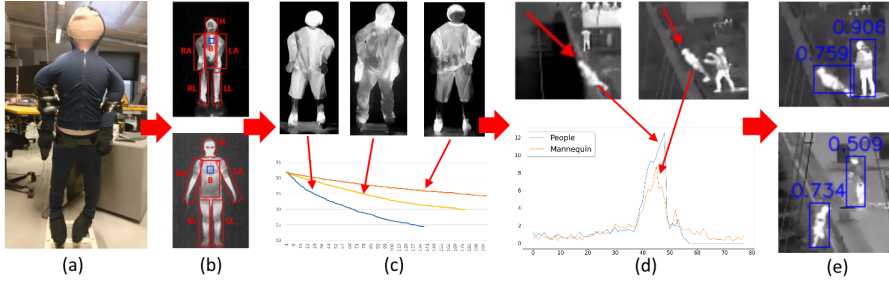


Fig. H.1: Overview of the proposed system and the investigation done in the paper. First, a human-like figure is designed with appropriate weights and support (a), followed by comparisons between the thermal signature of humans and the doll (b) and the stability of the signature through time (c). The fall motion of the figure is then compared to those of humans (d). Finally, the usability of the generated data is verified using a pedestrian detection algorithm (e).

To ensure that both categories of requirements are satisfied, a number of tests were conducted both in indoor laboratory settings as well as outdoor. We show that a simple inflated clothed doll can exhibit human-like temperature distribution in different body parts, such as the head, torso, hands and legs. Furthermore, this temperature distribution can be kept for a sufficiently long time in both warm and cold weather.

Fall data were also produced with the use of the doll in an outdoor environment and compared to the data produced by human volunteers. The motion of both the doll and the volunteer were estimated using optical flow, which showed minimum deviations between the two. Finally, the captured doll footage ran through a pedestrian detection algorithm and the results were compared to the ones from real human footage. The detection confidences of the two were very similar with only a 4% difference.

The main contributions of the paper revolve around proving the usability of an off-the-shelf doll for outdoor fall thermal data gathering. They can be summarized as:

- Experimentally verifying the visual representation of a doll construct for thermal imaging.
- Analyzing the motion representation of the doll for falling visualization.
- Presenting a real world use case of the captured thermal footage for pedestrian detection.

This work was created as a step-by-step guide to how a data-capturing system for emergency situations can be developed and verified. Its main use is for speeding up the initial process of capturing training and testing data for falling and drowning accidents, which normally require a great deal of overhead in both equipment and human-power. We show that the proposed doll system can be a viable alternative for data capture to other more complex and expensive systems [27–29] for the initial stages of building deep-learning-based fall detection systems.

The captured fall video dataset is provided as part of the publication—<https://www.kaggle.com/ivannikolov/thermal-mannequin-fall-image-dataset>.

2 Related Work

In this section, we discuss the related work in three main directions. Initially automated fall detection is discussed as a research area with different sensors and deep-learning techniques employed for different use cases, such as automated indoor and outdoor CCTV surveillance for anomaly detection, retirement homes and kindergarten surveillance, as well as medical facility for patient fall detection. All of these algorithms are shown to have the same requirements for a large amount of training and testing data.

This is why the next part of the related work discusses the ways mannequins are used for data capturing when it is hard to reproduce or dangerous scenarios are required. Finally, as we want to use thermal data for an uninterrupted 24-h fall detection, we focus on specifically thermal mannequins. This is required, as ensuring a mannequin gives off the same thermal signature as a real human is a non-trivial task that can significantly complicate the capturing setup.

2.1 Fall Detection

Fall detection systems can be roughly separated into indoor and outdoor use. The indoor use cases mostly focus on healthcare systems for children, the elderly or medical patients, while outdoor cases are connected to surveillance and accident prevention. Indoor fall detection systems can rely on using additional sensors and hardware for monitoring. These can be sensors mounted somewhere in the surroundings and used for surveying such as radar [30] or Dopplers [31], or sensors directly worn by people such as smart watches [32] or accelerometers [33].

These sensors can produce very precise time series data of a person’s position, motion and state but cannot be easily used in outdoor environments. Other possibilities are to use cameras for detection of falls. These

algorithms rely on classical approaches, such as Gaussian Mixture Models and contour-based template matching [34], pose estimation and SVMs [35] and optical flow oriented histogram analysis [36]. Deep-learning techniques, such as CNNs for activity prediction and skeletal pose extraction [16, 37], LSTM [1, 2] and autoencoders together with transfer learning and feature residuals [17, 38, 39], are also used on RGB, depth and thermal data.

Outdoor fall detection needs to be feasible in non-static environments with changing backgrounds, light conditions, density and clutter. This necessitates more robust approaches combining video motion detection and temporal features using three dimensional LSTM and ConvLSTM [12, 40], extracting depth from single images using Markov Random Fields and human detection using particle swarm optimization [41] or body posture analysis using two-branch multi-stage CNNs [42].

Outdoor fall detection can also be directed towards detecting falls from heights, such as ships using image patch clustering and HOG features [11] or falls in harbor fronts using optical flow [26] or convolutional autoencoders and YOLOv5 object detection [43]. All these algorithms require many examples of falls, which cannot be always captured easily and with a high level of reproductivity. Mannequins can be used in many of the cases to gather enough synthetic data of falls.

2.2 Mannequins in Data Capture

Using mannequins and dummies for gathering testing data has been an important part of the research and development in many fields. In the automotive industry, test dummies are used for gathering data from crashes to make vehicles safer [44, 45]. In medicine and the healthcare industry, mannequins are used for simulating emergency situations, such as falls of senior citizens [21] and medical patients [20, 46], through the use of both camera systems and wearable sensors.

Training doctors in performing diagnosis [47] and first responders on proper CPR techniques [48] are also areas where mannequin testing data are widely used. Mannequin torsos and heads are also regularly used in testing audio wave propagation and tuning devices [49, 50]. The flexibility of using mannequins provides the possibility to capture widely varying data for people's movements, poses and interactions.

Such data can either be immoral to capture with people [51] or require a considerable time investment [52, 53]. Mannequins are also used in robot vision where obstacle avoidance and interaction with humans are required, for generating RGB, depth or thermal images as well as point clouds [54, 55].

2.3 Thermal Mannequins

The use of mannequins in scenarios requiring the capture and evaluation of thermal data necessitates implementing specialized parts to simulate a thermal signature. Depending on the use case, either separate body parts, such as heads, torsos and feet, or full body male, female and child representations are produced [56]. All the thermal mannequins used in research contain complicated systems required to produce visuals and readings close to those of humans—multiple heat production zones [57], sweating [58, 59], soft tissue skin reactions [60], movement mechanisms [61] etc.

This becomes even more evident in for example work focusing on analyzing human thermoregulation using mannequins [61–63]. These mannequins are mostly suited for laboratory testing because of their size and required sensors and actuators connected to them. For outdoor environment testing, rescue mannequins and dummies can be used [27–29]; however, they still have the problems of high costs and being built with specific use cases. For generating thermal images of falling in water in outdoor environments, a simpler, lower-cost and easier to transport solution is required.

3 Capturing Cameras

Two cameras are used for the experiments in this paper. The experiments done on the harbor front use a Hikvision DS-2TD2235D-25 thermal camera [64], while all other experiments use an AXIS Q1921 thermal camera [65]. The Hikvision camera is pre-installed to monitor the harbor for safety by the city municipality, and the AXIS camera is used as a more mobile alternative as it provides internal parameters close to the Hikvision. Both are long-wavelength infrared (LWIR) cameras, which produce 8-bit grayscale images of relative temperature. The specifications of the two cameras are given in Table H.1.

Table H.1: The two thermal cameras used in the experiments. The main testing camera is the Hikvision DS-2TD2235D, while the AXIS Q1921 is used for laboratory testing. The Hikvision additionally contains an RGB sensor, which is not used in the experiments for this paper.

	AXIS Q1921	Hikvision DS-2TD2235D
Resolution	384 × 288	384 × 288
Image Sensor	Uncooled	Vanadium Oxide Uncooled
Response Waveband	8–13 μm	8–14 μm
Focal Length	19 mm	25 mm
Output	8-bit grayscale	8-bit grayscale
NETD	<100 milli-Kelvin	<50 milli-Kelvin

4. Thermal Doll Design

Both cameras capture video with a resolution of 288×384 and 25 frames per second and have comparable lens optics at 25 mm for the Hikvision camera, versus 19 mm for the AXIS camera. The Hikvision contains an additional RGB sensor, which is not used for the purposes of this paper. The Noise Equivalent Temperature Difference (NETD) specification of the two cameras differs with the Hikvision camera having a $\text{NETD} < 50$ milli-Kelvin, while the AXIS camera is rated at $\text{NETD} < 100$ milli-Kelvin.

The NETD is a measure of the size of difference between thermal points that the camera can distinguish, a smaller NETD specifying better contrast differentiation. As we use the AXIS camera for testing the thermal visualization, we speculate that, if the doll can be detected with the camera with the worse NETD rating, then it should be detected on the one with the better NETD. The Hikvision camera uses a vanadium oxide uncooled image sensor and has working wavelengths between 8 and $14 \mu\text{m}$, while the AXIS camera gives no specification for the image sensor, except that it is also uncooled and the working wavelengths were shown to be between 8 and $13 \mu\text{m}$ in [66].

4 Thermal Doll Design

As seen in Section 2.3, current state-of-the-art mannequins used for thermal data collection are expensive, hard to transport and heavy. This makes them not suitable for the use in outdoor tests, especially when the falls are into water. To address this, we selected a simple air-filled rubber doll as a basis of the design. This provides a human-like shape. The rubber exterior makes the process of drying off easier, and it can be easily inflated with an air compressor or pump. The height of the doll is 1.6 m, and its weight after being fully inflated is 1.5 kg.

As the doll will be thrown in a harbor, conventional heating solutions, such as electrical pads, vests, gel thermal pads etc. would be unusable, as the combination of sea water, dirt, seaweed and low temperature would easily degrade and destroy them. A simpler solution was thus selected where water is boiled and put in sealed thermoses. Before each experiment, the water is poured onto the doll.

As there will not be a solution to continuously provide heat to the doll's exterior, and it is made out of rubber which has bad thermal conductivity and retention, a layer of clothes is required. Each part of the doll's body that should be detected by the thermal cameras is clothed, using polyester clothing consisting of a tracksuit, a sweatshirt with a hood, gloves, socks and a winter hat. The clothes are chosen as the material would keep the heat from the applied water, without losing their shape or shrinking.

To make the doll behave closer to a human when thrown, four training ankle weights are strapped to it—one on each hand and leg. Each weight is 2

kg, boosting the overall weight of the mannequin to 9.5 kg. Finally, because of the additional weight strapped to it, together with the weight from the wet clothes, the mannequin requires additional structural support. Aluminum tube supports are made for each leg and connected to a main structure at the lower back of the doll. A heavy base stand is also made, with slots for the leg supports, so the doll can be set up standing for the easier pouring of hot water onto it. The final construction can be seen in Fig. H.2.



Fig. H.2: View of the created low-cost doll, together with clothes and ankle weights.

5 Doll Thermal Appearance

To test the thermal visuals of the doll after hot water is poured over it, two experiments were conducted. The first one aimed to compare the temperature of different body parts of the doll to the temperature of humans in the same environment. This would show that the visual representation of the clothed doll would be close enough to a human, when thrown in the water. The second experiment would test the temperature change of the clothed doll over time. This is necessary as no persistent source of heat is applied, and it is expected that the initial heat from the hot water would dissipate over time, especially in colder weather. For both experiments, we used the AXIS Q1921 thermal camera.

5.1 Comparing Temperature between the Doll and Real People

The first test was designed to determine if the clothed doll would exhibit temperature readings similar to those seen in real humans after the clothes were positioned on it and the hot water has been poured on it. A static laboratory environment was chosen for this test so that any possible effects of environmental variables can be minimized. For this test, we took inspiration from the comparison between human and mannequin thermal visuals presented in the research by [63, 67], together with the separation of the body in heat zones for the detection of different facial expressions presented in the work by [68].

We captured the thermal visuals of 11 volunteers, as well as the doll from four distances—1, 2, 3 and 4 m. We chose the nearer distance as only a very small part of the participants could be seen at the farther distance due to the constraints of the laboratory. We separated six thermal zones—head (H), body (B), left (LA) and right (RA) arm and left (LL) and right (RL) leg. Example of these zones on a participant and the doll can be seen in Fig. H.3.

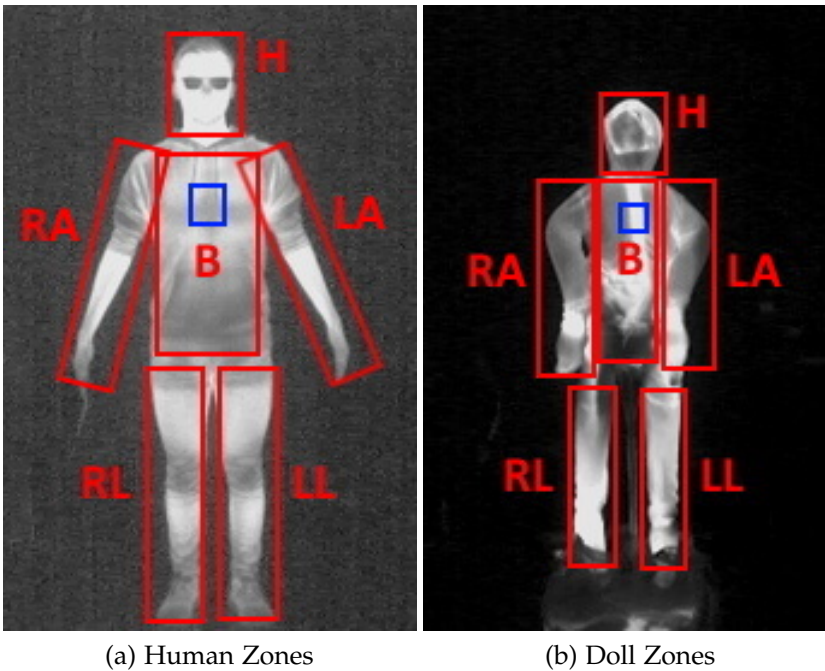


Fig. H.3: Visualization of the six thermal zones—head (H), body (B), left arm (LA), right arm (RA), left leg (LL) and right leg (RL) on one of the volunteers and the doll. Blue squares show the central pixels used for calculating the thermal image calibration.

The AXIS camera captures uncalibrated thermal images which change their intensity depending on the real maximum temperature of the captured object and surrounding environment. This can cause variations in the thermal images. To minimize these variations, a simple scaling step was included so that all captured intensities were transformed to real world temperatures.

To do this, we used an infrared thermometer to measure the temperature in Celsius in the body center of each participant and the doll. We then calculated the average intensity of a square of pixels from the captured thermal images in the center of the body of each participant, and the doll at each captured distance (seen as blue squares in Fig. H.3).

The ratio between the thermometer measurements and the average intensities can then be used to scale the intensities of all other pixels of each image to real world temperatures. Fig. H.4 shows the average temperature value for each body part of the participants compared to the doll's body part temperature. The doll had an overall higher temperature in the leg zones and lower temperature in the head zone. The readings were relatively stable between the three distance measurements. The results show that the clothed doll exhibited overall temperature readings close to the ones from humans and that it can be used as a visual replacement of a human.

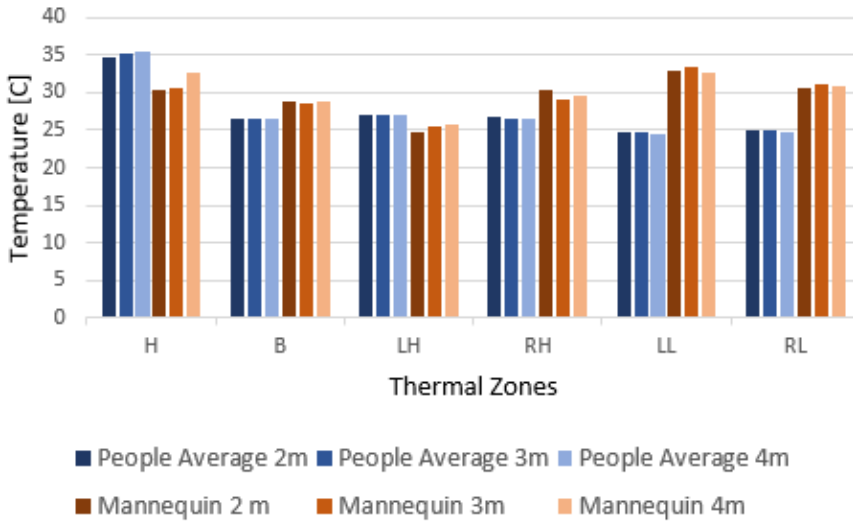


Fig. H.4: Average temperatures of the volunteers' thermal zones, compared to the ones from the doll captured from 4, 3 and 2 m, respectively.

5.2 Doll Temperature Change over Time

It was shown that the doll's temperature representation after hot water has been poured over it closely resembles the ones from real humans. The change over time of the doll's temperature needs to be studied, as it is not dynamically maintained over time. It is expected that, without a constant source of heat, the initial measured temperature would decline steadily and that the speed of the decline would depend on the environment in which it is. Knowing how the temperature would change over time is necessary so that enough time is given for performing the fall experiments and generating data while the temperature remains optimal.

We measured the change of the doll's temperature from the time the hot water was poured until the difference between it and the background becomes small enough that it would be hard to distinguish the doll. In our case, this difference should be at least 5 °C. We measured this change in three different scenarios—outdoor in cold weather, outdoor in mild weather and indoor. For the temperature readings, we calculated the average temperature in Celsius using the same scaling presented in the previous section. The results from these measurements are given in Fig. H.5.

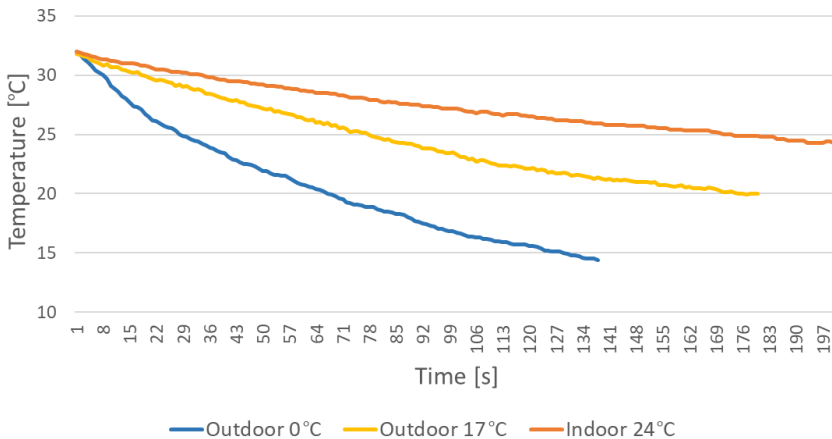


Fig. H.5: Measured doll temperature over time in three different scenarios—outdoor in cold weather at 0 °C, outdoor in mild weather at 17 °C and indoor at 24 °C.

The cold weather scenario was captured at an environmental temperature of 0 °C. The warm weather scenario was captured at an environmental temperature of 17 °C and the indoor scenario at a temperature of 24 °C. In both outdoor scenarios, there was wind present. The temperature of the doll changed at an average rate of 0.12 °C/s in cold weather, at 0.06 °C/s in mild weather and at 0.03 °C/s indoors. In the outdoor scenarios, the wind lowered the temperature faster, with the cold weather contributing to

an even faster decline. On the other hand, the difference between the environment and the doll was much larger in cold weather, compared with that in the warmer weather and indoors. In all cases, there should be at least 3 to 4 mins for performing a fall. This would require a re-application of hot water after each fall.

6 Fall Motion Comparison

To obtain better insights into the fall behavior of the created doll construction, it was compared to real people falling in the harbor front. For this test, the Hikvision DS-2TD2235D camera was used, mounted in the outdoor environment where it will be used—overlooking a harbor front in the city of Aalborg, Denmark, as seen in Fig. H.6.

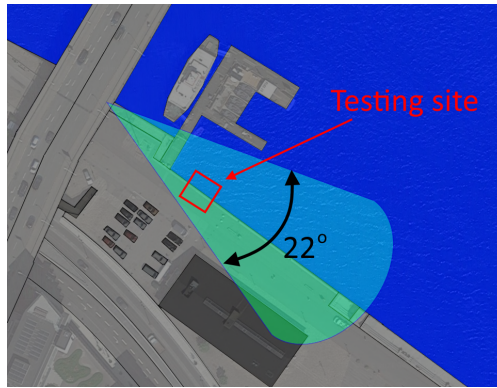


Fig. H.6: The area and field of view of the Hikvision DS-2TD2235D camera used for the outdoor experiments.

We used five videos captured as part of the publication by [26] of five real people falling in different ways—from stationary position, while walking and while running. We then captured videos of the doll in the same place. All in all, 22 videos of falls were captured in different scenarios—being moved along the edge and pushed in, falling in without exterior help, being thrown in, being kicked in, etc. Example frames from the people and doll tests can be seen in Fig. H.7(a) and (b), with the volunteer and the doll shown with a red arrow.

First, the number of frames from the person starting falling to hitting the water was measured. The average number of frames of the recorded participants was 23. This was compared with the average number of frames of the doll falls, which in our case was 29 frames. The longer fall time can be explained as the doll was pushed close to the harbor wall, while the volunteers were required to jump farther away from the wall for safety reasons.

6. Fall Motion Comparison

The doll was also affected more by the strong wind than the volunteers, because of its weight. To compare the motion of falling for the doll and the volunteers, optical flow was used as it is widely seen in the literature.

We first calculated the optical flow from the videos and calculated the maximum vector magnitude in a square area on the edge of the harbor, where volunteers and the doll were falling for each frame. This area was manually selected and annotated to minimize the change of errors. This gave us a time-wise signal of the maximum detected position change and simplified the comparison, by eliminating variations in their orientation. Examples of the changes of the optical flow magnitude for the five volunteers can be seen in Fig. H.7(c) and for the first five doll experiments in Fig. H.7(d). Clear peaks can be seen with a cutoff when hitting the water, showing that a fall can be detected in both cases.

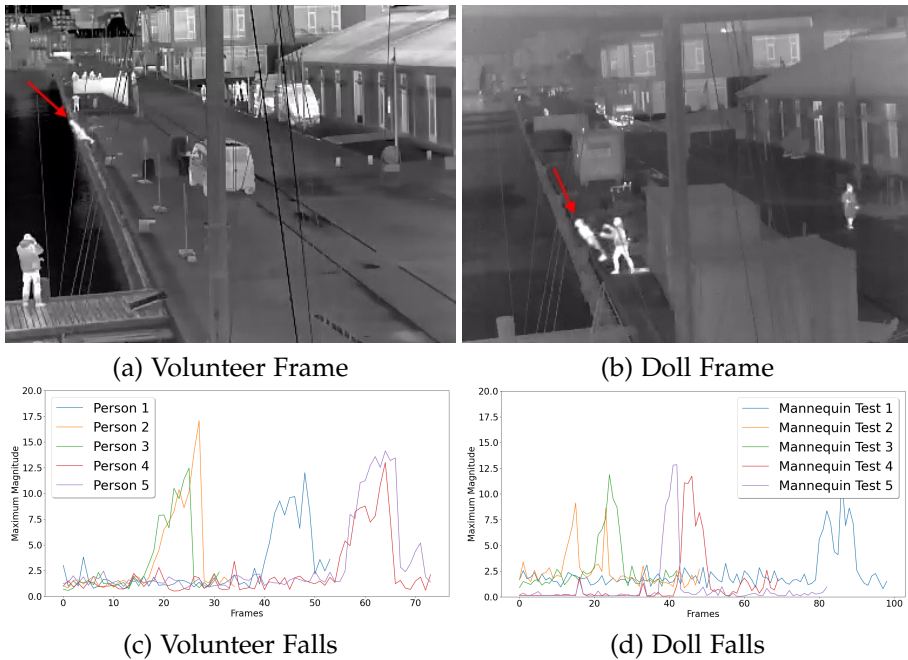


Fig. H.7: Examples of frames from a volunteer (a) and a doll (b) falling, shown with a red arrow, together with the maximum magnitudes of the optical flow vectors of five videos from the falling person and the doll (c)(d). The clear spike when they fall can be seen. The spike is in different positions as it took different durations of time before the fall.

For an easier comparison, the peaks were detected, and the signals were registered so that all the peaks overlap. To compare the volunteer and the doll results, we padded them to have the same length and calculated their average. The mean signal for each can be visually compared in Fig. H.8. The maximum peaks of each, which were 12.54 pixels/frame for the people

and 10.48 pixels/frame for the doll show that the captured movement was comparable between the two.

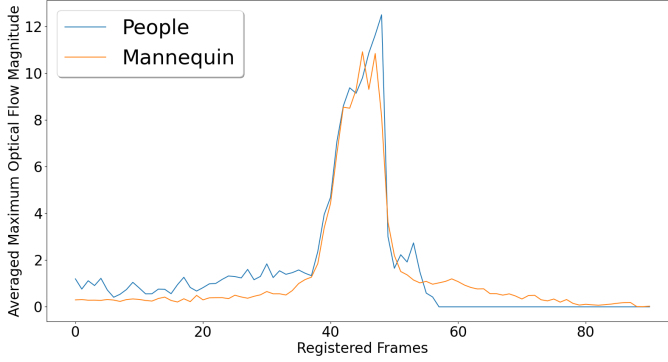


Fig. H.8: The averaged maximum optical flow magnitudes of the different people and the doll falls. Before averaging the fall data, the peaks were registered.

To further compare the captured optical flow magnitude signals, we used dynamic time warping (DTW) [69]. The technique is useful for comparing time series signals which are not perfectly aligned and with different lengths, when simply calculating an Euclidean distance between them would not work. DTW can be used in this case as both magnitude signals have the same indexing and the same sampling. Dynamic time warping calculates the difference between the current, previous and next points in both signals and uses these as costs, selecting the minimum ones. In this way, even if the two signals change with different speeds, the correct indices can be used for comparing them.

As we did not have a ground truth, we used the average volunteer optical flow time signal as one. We calculated the DTW distance between the average volunteer ground truth and each of the 22 doll optical flow signals. For comparison, we also calculated the DTW distance from the average ground truth to each of the volunteer signals. This will give an idea of how the doll fall behavior compares to variations of the captured volunteer fall behavior.

The average DTW distance between the ground truth and the doll tests was 60.16, while the average DTW between the ground truth and the volunteer tests was 33.84. This shows that, even though the peak values of the volunteer and the doll falls were quite similar, the overall trajectories differed. This can also be explained with the difference in weight and the safety requirement that the participants jumped farther away from the harbor wall, while the doll was both thrown farther and pushed close to the wall.

7 Doll Detection Comparison

To show how the doll fall videos compare to real human ones from the point of view of a real computer vision application, we used an object detector trained to detect pedestrians. For this, we chose the YOLOv5 model [70] as it provides state-of-the-art performance and has been proven robust on thermal data [71, 72]. The model was trained on data from the multi-seasonal LTD dataset [73] so that it had the best possible chance of detection. The model was trained for 200 epochs with a learning rate of 0.00075 on a NVIDIA RTX2080Ti graphics card.

Both the volunteer and the doll videos were separated into frames, and only the ones between jumping off the harbor edge and hitting the water were selected. This was done to limit the test to only specific instances connected to falling in the water and the behavior leading to the fall. The frames after hitting the water were also skipped. In the case of the volunteer videos, people stayed mostly submerged and then swam out of the field of view of the camera. The doll lost heat very fast after hitting the cold water, making it hard to distinguish the doll from the background.

The detection confidence score for the doll and the volunteers for each frame was saved, and an average confidence score was calculated for each of the videos. From these scores, an overall average score was calculated for all the doll and the volunteer videos. In addition, the percentage of frames in which the volunteers and the doll were detected in each video was also calculated. The third calculated value shows how many frames the object detector loses track before the person or the doll hits the water. This is important as both the volunteers and the doll change their orientation and shape in the air by bending their limbs, thus, making it harder for YOLOv5 to detect them.

All the three are given in Table H.2. The average detection confidence scores of the volunteers and the doll videos are very similar as well as the number of detected frames. On the other hand, in 3 of the 22 doll videos, the YOLOv5 model could not detect the doll. Upon further inspection, two of these videos show the doll being thrown at an angle that is close to horizontal to the camera view, while in the third video the doll has not been sufficiently warmed up, making it difficult to distinguish. Problem frames with a horizontal view and the doll not warm enough are shown in Fig. H.9(a) and (b).

Some interesting observations can be made from the last row of Table H.2. For the doll videos, the model loses track of an average of two frames before the doll hits the water and a maximum of four frames. For the volunteer videos, the model loses track of an average of one frame before the volunteers hit the water and a maximum of two. This can be attributed to the fact that once the doll was thrown or pushed, its hands and legs moved very little

Table H.2: Average YOLOv5 detection results for the volunteers and the doll videos. These consist of confidence scores for the detected instances in the frames, percentage of frames in which the volunteer or the doll have been detected and the number of frames before hitting the water where the volunteers or the doll were not detected.

Results	Volunteers	Doll
Avg. Confidence Score	0.761	0.730
Avg. Detected Frames [%]	77.7	62.1
Avg. Lost Frames before Water Hit	1	2

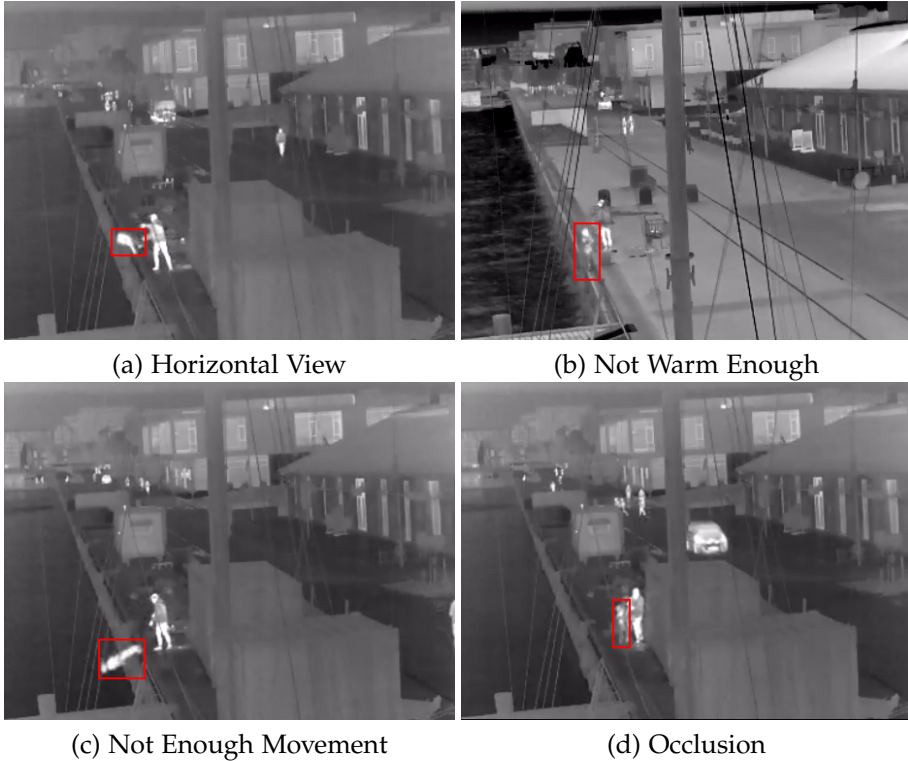


Fig. H.9: Problems seen with the doll clips. The object detection can fail before the doll is thrown if it is not warm enough (b) or is occluded by a person or object (d). After the doll is pushed or thrown, the object detection can fail if the doll body is horizontal and cannot be seen by the camera (a) or not enough movement is present in the limbs (c).

even with the added weights, making the overall shape less human-like (Fig. H.9(c)).

Finally, the lower percentage of frames that the doll was detected compared to the volunteers can be explained with the fact that the doll was carried or propped up on objects before falling in the water, which can result

in occlusions and failed detection (Fig. H.9(d)). Examples of successfully detected doll frames from different clips, together with confidence scores can be seen in Fig. H.10. It can be seen that the doll was detected with similar confidences as the other people present in the image.

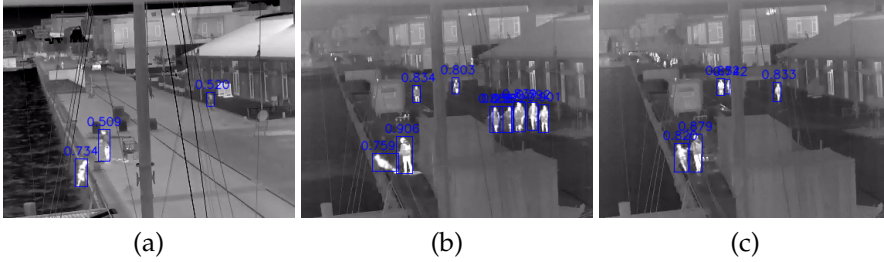


Fig. H.10: Examples of successful doll detection from the YOLOv5 model together with the confidence scores. In both scenarios where there are few people (a) and many other people (b)(c), the doll is detected before it hits the water.

The major strength of the proposed work is the straightforward and easily replicable pipeline for developing and verification. We show that, even though the physical measurements of the doll are not directly comparable to a human, the captured thermal data and movement trajectories are sufficiently convincing that they can be used as the data for a deep-learning system. The proposed doll can be also easily deployed in “in-the-wild” scenarios, with minimal logistical overhead. Having demonstrated that thermal data can be captured from the doll extends its usability beyond only RGB data capture that most of the reviewed related work was using mannequins for.

8 Conclusions and Future Work

Detecting falls into water is an important step for preventing drowning accidents. Drowning is a major public health problem that can occur at any point of time. This makes it necessary for automated surveillance to be able to detect and signal accidents as soon as they happen. Footage of fall accidents like these is captured rarely, making it necessary to synthetically generate enough diverse data for the successful training and testing of automatic systems. Involving real people can pose a health and safety risk, and this brings ethical concerns about preserving anonymity. On the other hand, generating data using simulation and deep-learning methods can provide imperfect results with the necessity of post-processing steps.

This is why, in this paper, we demonstrated that a rubber air-filled doll can be used for generating thermal image fall data in outdoor scenarios. By using

off-the-shelf clothes and hot water, the doll can achieve human-like thermal properties and maintain them for an extended period of time in different weather conditions. Throwing and pushing the doll off a ledge also achieved similar movement vectors to real people jumping.

Finally, we showed that a YOLOv5 object detection algorithm trained on people can detect the thermal signature of the doll with confidence close to the one for detecting real people. As this is the first dataset focused on generated thermal fall data and the detection of such accidents is crucial for automatic surveillance systems, we made the dataset available online so that others can benefit from our data.

The proposed solution can be also applicable to other domains, such as generating traffic accident data, as well as indoor anomaly detection use cases, such as generating fall data for the elderly and children.

We encountered certain limitations in the proposed work. The rubber doll, even with the added supports and weights, had an overall weight of 9.5 kg, which is far from a regular human weight. This helped with making the system more easily transportable but lowered the precision of the captured fall movement data. The simple method of raising the temperature of the doll can also be viewed as a positive feature and a limitation. The inability to maintain a stable temperature for long periods of time would require additional equipment for performing repeated experiments that need precise temperatures.

Potential improvements for the proposed solution can be done by addressing the difference in the weight and articulation between the doll and a real human. To address this, adding a weighted vest is proposed, together with 3D printed joints for the arms and legs of the doll. This would give the possibility to add more weight—up to 15 more kilograms, making the full weight of the doll up to 25 kg and matching other off-the-shelf doll solutions but with the added benefit of flexibility. This would also make the additional burden on transportation and setup less problematic.

For maintaining the thermal signature of the doll across time, we propose adding heating thermal pads to the clothing. These pads would be connected to isolated thermal sensor pads and an Arduino or Raspberry-Pi in a water-tight case. In this way, data of the clothes temperature of the doll can be sent through a low-power Bluetooth or radio signal to a monitoring station, and when necessary the pads can be changed or more heating can be applied.

References

- [1] Raunak Manekar, Sumeet Saurav, Somsukla Maiti, Sanjay Singh, Santanu Chaudhury, Ravi Kumar, Kamal Chaudhary, et al., “Activity recognition for indoor fall detection in 360-degree videos using deep learning

- techniques,” in *Proceedings of 3rd International Conference on Computer Vision and Image Processing*. Springer, 2020, pp. 417–429.
- [2] Hamidreza Sadreazami, Miodrag Bolic, and Sreeraman Rajan, “Tl-fall: contactless indoor fall detection using transfer learning from a pre-trained model,” in *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2019, pp. 1–5.
 - [3] Waqas Sultani, Chen Chen, and Mubarak Shah, “Real-world anomaly detection in surveillance videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.
 - [4] Raghavendra Chalapathy and Sanjay Chawla, “Deep learning for anomaly detection: A survey,” *arXiv preprint arXiv:1901.03407*, 2019.
 - [5] Huansheng Song, Haoxiang Liang, Huaiyu Li, Zhe Dai, and Xu Yun, “Vision-based vehicle detection and counting system using deep learning in highway scenes,” *European Transport Research Review*, vol. 11, no. 1, pp. 1–16, 2019.
 - [6] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, and Partha Pratim Roy, “Anomaly detection in road traffic using visual surveillance: A survey,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–26, 2020.
 - [7] Yole Développement, “Thermal imagers and detectors 2020 - covid-19 outbreak impact – preliminary report,” http://www.yole.fr/Thermal_Imagers_And_Detectors_Covid19_Outbreak_Impact.aspx, 2020, last accessed: August, 2021.
 - [8] Allied Market Research, “Global thermal imaging camera market by 2030,” <https://www.globenewswire.com/news-release/2021/08/09/2277188/0/en/Global-Thermal-Imaging-Camera-Market-is-Expected-to-Reach-7-49-Billion-by-2030-Says-AMR.html>, 2021, last accessed: August, 2021.
 - [9] Nirmalya Thakur and Chia Y Han, “Country-specific interests towards fall detection from 2004–2021: An open access dataset and research questions,” *Data*, vol. 6, no. 8, pp. 92, 2021.
 - [10] Muhammad Mubashir, Ling Shao, and Luke Seed, “A survey on fall detection: Principles and approaches,” *Neurocomputing*, vol. 100, pp. 144–152, 2013.
 - [11] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Dimitris Dres, and Dimitris Drakoulis, “Man overboard event detection from rgb and thermal imagery: possibilities and limitations,” in *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, 2020, pp. 1–6.

- [12] Na Lu, Yidan Wu, Li Feng, and Jinbo Song, "Deep learning for fall detection: Three-dimensional cnn combined with lstm on video kinematic data," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 314–323, 2018.
- [13] Jesús Gutiérrez, Víctor Rodríguez, and Sergio Martin, "Comprehensive review of vision-based fall detection systems," *Sensors*, vol. 21, no. 3, pp. 947, 2021.
- [14] Yong Chen, Weitong Li, Lu Wang, Jiajia Hu, and Mingbin Ye, "Vision-based fall event detection in complex background using attention guided bi-directional lstm," *IEEE Access*, vol. 8, pp. 161337–161348, 2020.
- [15] Qi Feng, Chenqiang Gao, Lan Wang, Yue Zhao, Tiecheng Song, and Qiang Li, "Spatio-temporal fall event detection in complex scenes using attention guided lstm," *Pattern Recognition Letters*, vol. 130, pp. 242–249, 2020.
- [16] Tsung-Han Tsai and Chin-Wei Hsu, "Implementation of fall detection system based on 3d skeleton for deep learning technique," *IEEE Access*, vol. 7, pp. 153049–153059, 2019.
- [17] Jacob Nogas, Shehroz S Khan, and Alex Mihailidis, "Fall detection from thermal camera using convolutional lstm autoencoder," in *Proceedings of the 2nd workshop on Aging, Rehabilitation and Independent Assisted Living, IJCAI Workshop*, 2018.
- [18] Umar Asif, Stefan Von Cavallar, Jianbin Tang, and Stefan Harrer, "Sshfd: Single shot human fall detection with occluded joints resilience," *arXiv preprint arXiv:2004.00797*, 2020.
- [19] Umar Asif, Benjamin Mashford, Stefan Von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, and Stefan Harrer, "Privacy preserving human fall detection using video data," in *Machine Learning for Health Workshop*. PMLR, 2020, pp. 39–51.
- [20] Mary E Bowen, Jeffrey Craighead, Chadwick A Wingrave, and William D Kearns, "Real-time locating systems (rtls) to improve fall detection," *Gerontechnology*, vol. 9, no. 4, pp. 464, 2010.
- [21] Yasar Guneri Sahin, AA Eren, Ahmet Reha Seker, and E Okur, "A personalized fall detection system for older people," in *Rhodes (Greece): Proceedings of the 2013 international conference on biology and biomedicine*, 2013, pp. 43–48.
- [22] World Heath Organization (WHO), "Drowning accidents report," <https://www.who.int/news-room/fact-sheets/detail/drowning>, 2021, last accessed: November, 2021.

- [23] Juan Du, "Characteristics and function analysis of swimming life saving system based on machine vision technology," in *Journal of Physics: Conference Series*. IOP Publishing, 2021, vol. 1881, p. 042079.
- [24] Abdel Ilah N Alshbatat, Shamma Alhameli, Shamsa Almazrouei, Salama Alhameli, and Wadhha Almarar, "Automated vision-based surveillance system to detect drowning incidents in swimming pools," in *2020 Advances in Science and Engineering Technology International Conferences (ASET)*. IEEE, pp. 1–5.
- [25] Jisoo Park, Jingdao Chen, Yong K Cho, Dae Y Kang, and Byung J Son, "Cnn-based person detection using infrared images for night-time intrusion warning systems," *Sensors*, vol. 20, no. 1, pp. 34, 2020.
- [26] Soren Bonderup, Jonas Olsson, Morten Bonderup, and Thomas B Moeslund, "Preventing drowning accidents using thermal cameras," in *International Symposium on Visual Computing*. Springer, 2016, pp. 111–122.
- [27] Simulaids, "Ti rescue randy," <https://www.aedsuperstore.com/simulaids-ti-rescue-randy-thermal-imaging-mankin.html>, last accessed: September, 2021.
- [28] Thermetrics, "Newton thermal manikin," <https://thermetrics.com/products/manikin/newton-thermal-manikin/>, last accessed: September, 2021.
- [29] Lion, "Smartdummy thermal manikin," <https://www.lionprotects.com/smartdummy-thermal-manikin>, last accessed: September, 2021.
- [30] Abhijit Bhattacharya and Rodney Vaughan, "Deep learning radar design for breathing and fall detection," *IEEE Sensors Journal*, vol. 20, no. 9, pp. 5072–5085, 2020.
- [31] Branka Jokanović and Moeness Amin, "Fall detection using deep learning in range-doppler radars," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 1, pp. 180–189, 2017.
- [32] Taylor R Mauldin, Marc E Canby, Vangelis Metsis, Anne HH Ngu, and Coralys Cubero Rivera, "Smartfall: A smartwatch-based fall detection system using deep learning," *Sensors*, vol. 18, no. 10, pp. 3363, 2018.
- [33] Guto Leoni Santos, Patricia Takako Endo, Kayo Henrique de Carvalho Monteiro, Elisson da Silva Rocha, Ivanovitch Silva, and Theo Lynn, "Accelerometer-based human fall detection using convolutional neural networks," *Sensors*, vol. 19, no. 7, pp. 1644, 2019.

- [34] Subhash Chand Agrawal, Rajesh Kumar Tripathi, and Anand Singh Jalal, "Human-fall detection from an indoor video surveillance," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2017, pp. 1–5.
- [35] Zhanyuan Huang, Yang Liu, Yajun Fang, and Berthold KP Horn, "Video-based fall detection for seniors with human pose estimation," in *2018 4th International Conference on Universal Village (UV)*. IEEE, 2018, pp. 1–4.
- [36] Tian Wang and Hichem Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988–998, 2014.
- [37] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 2017, pp. 81–84.
- [38] Faten A Elshwemy, Reda Elbasiony, and Mohamed Talaat Saidahmed, "A new approach for thermal vision based fall detection using residual autoencoder," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 2, pp. 250–258, 2020.
- [39] Fouzi Harrou, Nabil Zerrouki, Ying Sun, and Amrane Houacine, "An integrated vision-based approach for efficient human fall detection in a home environment," *IEEE Access*, vol. 7, pp. 114966–114974, 2019.
- [40] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding, "Anomaly detection in video sequences: A benchmark and computational model," *arXiv preprint arXiv:2106.08570*, 2021.
- [41] Myeongseob Ko, Suneung Kim, Mingi Kim, and Kwangtaek Kim, "A novel approach for outdoor fall detection using multidimensional features from a single camera," *Applied Sciences*, vol. 8, no. 6, pp. 984, 2018.
- [42] Jin Zhang, Cheng Wu, and Yiming Wang, "Human fall detection based on body posture spatio-temporal evolution," *Sensors*, vol. 20, no. 3, pp. 946, 2020.
- [43] Jinsong Liu, Mark P Philipsen, and Thomas B Moeslund, "Supervised versus self-supervised assistant for surveillance of harbor fronts," pp. 610–617, 2021.
- [44] Tao Xu, Xiaoming Sheng, Tianyi Zhang, Huan Liu, Xiao Liang, and Ao Ding, "Development and validation of dummies and human models used in crash test," *Applied bionics and biomechanics*, vol. 2018, 2018.

- [45] Mohamed Karim Belaid, Maximilian Rabus, and Ralf Krestel, "Crashnet: an encoder–decoder architecture to predict crash test outcomes," *Data Mining and Knowledge Discovery*, pp. 1–22, 2021.
- [46] Gina E Bertocci, Mary Clyde Pierce, Ernest Deemer, Fernando Aguel, Janine E Janosky, and Ev Vogeley, "Using test dummy experiments to investigate pediatric injury risk in simulated short-distance falls," *Archives of pediatrics & adolescent medicine*, vol. 157, no. 5, pp. 480–486, 2003.
- [47] Roberto Martinez-Maldonado, Tamara Power, Carolyn Hayes, Adrian Abdiprano, Tony Vo, Carmen Axisa, and Simon Buckingham Shum, "Analytics meet patient manikins: Challenges in an authentic small-group healthcare simulation classroom," in *Proceedings of the seventh international learning analytics & knowledge conference*, 2017, pp. 90–94.
- [48] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler, "Detecting mistakes in cpr training with multimodal data and neural networks," *Sensors*, vol. 19, no. 14, pp. 3099, 2019.
- [49] Andrea Genovese and Agnieszka Roginska, "Hmdir: An hrtf dataset measured on a mannequin wearing xr devices," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.
- [50] Simone Spagnol, Kristján Bjarki Purkhús, Rúnar Unnthórsson, and Sverrir Karl Björnsson, "The viking hrtf dataset," in *16th Sound and music computing conference*. Sound and Music Computing Network, 2019, pp. 55–60.
- [51] Muhammad Alhammami, Chee-Pun Ooi, and Wooi-Haw Tan, "Violent actions against children," *Data in brief*, vol. 12, pp. 480–484, 2017.
- [52] Shuangjun Liu, Yu Yin, and Sarah Ostadabbas, "In-bed pose estimation: Deep learning with shallow dataset," *IEEE journal of translational engineering in health and medicine*, vol. 7, pp. 1–12, 2019.
- [53] Guy Satat, Matthew Tancik, Otkrist Gupta, Barmak Heshmat, and Ramesh Raskar, "Object classification through scattering media with deep learning on time resolved measurement," *Optics express*, vol. 25, no. 15, pp. 17466–17479, 2017.
- [54] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft, "Deepanomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field," *Sensors*, vol. 16, no. 11, pp. 1904, 2016.

- [55] Mikkel Fly Kragh, Peter Christiansen, Morten Stigaard Laursen, Morten Larsen, Kim Arild Steen, Ole Green, Henrik Karstoft, and Rasmus Nyholm Jørgensen, "Fieldsafe: dataset for obstacle detection in agriculture," *Sensors*, vol. 17, no. 11, pp. 2579, 2017.
- [56] Yehu Lu, Kalev Kuklane, and Chuansi Gao, "Types of thermal manikin," in *Manikins for textile evaluation*, pp. 25–54. Elsevier, 2017.
- [57] Robert B Farrington, John P Rugh, Desikan Bharathan, and Rick Burke, "Use of a thermal manikin to evaluate human thermoregulatory responses in transient, non-uniform, thermal environments," *SAE transactions*, pp. 548–556, 2004.
- [58] J Fan, "Recent developments and applications of sweating fabric and applications of sweating fabric manikin—walter," *Thermal manikins and modelling*, pp. 202–209, 2006.
- [59] Y Chen, J Xu, and J Fan, "Passive and active water supply to perspiring manikin," in *Proceedings of 6th International Thermal Manikin and Modeling Meeting*, 2006, pp. 16–18.
- [60] W Yu, J Fan, SP Ng, and HB Gu, "Female torso mannequins with skeleton and soft tissue for clothing pressure evaluation," in *Thermal manikins and modeling. Sixth international thermal manikin and modeling meeting (613M)*. Hong Kong: The Hong Kong Polytechnic University, 2006.
- [61] Håkan O Nilsson, *Comfort climate evaluation with thermal manikin methods and computer simulation models*, Ph.D. thesis, Bygghälsan, 2004.
- [62] Jun Miura, Mitsuhiro Demura, Kaichiro Nishi, and Shuji Oishi, "Thermal comfort measurement using thermal-depth images for robotic monitoring," *Pattern Recognition Letters*, vol. 137, pp. 108–113, 2020.
- [63] Shaun Fitzgerald, Henry Atkins, Ryan Leknys, and Richard Kelso, "A thermal test system for helmet cooling studies," in *Multidisciplinary Digital Publishing Institute Proceedings*, 2018, vol. 2, p. 272.
- [64] Hikvision, "Ds-2td2235d-25/50," <https://us.hikvision.com/en/products/more-products/discontinued-products/thermal-camera/thermal-network-bullet-camera-ds>, 2015, last accessed: September, 2021.
- [65] AXIS, "Q1921," <https://www.axis.com/en-us/products/axis-q1921-e>, 2018, last accessed: September, 2021.
- [66] Norbert Schuster and John Franks, "Depth of field in modern thermal imaging," in *Infrared Imaging Systems: Design, Analysis, Modeling, and*

References

- Testing XXVI*. International Society for Optics and Photonics, 2015, vol. 9452, p. 94520J.
- [67] Akinaru Iino, Tetsuo Annaka, Yukari Iino, and Masaaki Ohba, "Visualization of sensible heat on thermal mannequin's surface by image analysis of infrared animation," in *The Fourth International Conference on Advances in Wind and Structures (AWAS2008)*, 2008.
- [68] Braj Bhushan, Sabnam Basu, Pradipta Kumar Panigrahi, and Sourav Dutta, "Exploring the thermal signature of guilt, shame, and remorse," *Frontiers in Psychology*, vol. 11, pp. 2874, 2020.
- [69] Stan Salvador and Philip Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
- [70] Ultralytics, "Yolov5," <https://github.com/ultralytics/yolov5>, 2020, last accessed: October, 2021.
- [71] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar, "Thermal object detection in difficult weather conditions using yolo," *IEEE Access*, vol. 8, pp. 125459–125476, 2020.
- [72] Noor Ul Huda, Bolette D Hansen, Rikke Gade, and Thomas B Moeslund, "The effect of a diverse dataset for transfer learning in thermal person detection," *Sensors*, vol. 20, no. 7, pp. 1982, 2020.
- [73] Ivan Adriyanov Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Velling Dueholm, Anders Skaarup Johansen, Kamal Nasrollahi, and Thomas B Moeslund, "Seasons in drift: A long-term thermal imaging dataset for studying concept drift," in *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-872-4

AALBORG UNIVERSITY PRESS