Facial Emotion Recognition for Citizens with Traumatic Brain Injury for Therapeutic Robot Interaction

Ilyas, Chaudhary Muhammad

DOI (link to publication from Publisher):
10.54337/aau460118975

Publication date:
2021

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

FACIAL EMOTION RECOGNITION FOR CITIZENS WITH TRAUMATIC BRAIN INJURY FOR THERAPEUTIC ROBOT INTERACTION

BY
CHAUDHARY MUHAMMAD AQDUS ILYAS

DISSERTATION SUBMITTED 2021

AALBORG UNIVERSITY
DENMARK

Facial Emotion Recognition for Citizens with Traumatic Brain Injury for Therapeutic Robot Interaction

Ph.D. Dissertation

Chaudhary Muhammad Aqdus Ilyas

Dissertation submitted June 30, 2021

Dissertation submitted: September 2021

PhD supervisors: Dr. Matthias Rehm
Aalborg University

Professor Kamal Nasrollahi
Aalborg University

PhD committee: Associate Professor David Meredith (chairman)
Aalborg University

Professor Britta Wrede
Bielefeld University

Professor Dirk Heylen
Twente University

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

Department: Department of Architecture,
Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-991-6

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

Printed in Denmark by Rosendahls, 2021

Curriculum Vitae

Chaudhary Muhammad Aqdus Ilyas

Chaudhary Muhammad Aqdus Ilyas received his Master's degree in Electrical and Electronics Engineering from Coventry University, UK in 2016. He worked as research assistant at the Coventry University, with the department of Aviation, Aerospace, Electrical and Electronic Engineering. From July 2011 to September 2012, Aqdus Ilyas worked as Embedded System Software Engineer at NexTech Services, Islamabad, main tasks involved developing software to embedded systems. During the PhD studies he spent four months research stay at the University of Western Australia, Perth, Australia. His main interests are computer vision and robotics, particularly in the area of robotic rehabilitation of people suffering with brain injuries through the use of computer vision techniques. He has been involved in the supervision of graduate and postgraduate students in the area of image processing, computer vision and robotics. Currently, he is working in the University of Cambridge, UK, as a Research Assistant/Associate and working on a project for affective wellbeing of workers in a working environment.

Abstract

Humans are capable of producing a large number of facial expressions (FE) by the activation of facial muscles. Facial expression recognition (FER) is active research area where human emotions are determined by the classification of facial muscle movements. Automatic recognition of facial emotions is of prime importance for wide range of applications such as bio-metric security and surveillance, human-machine or human-robot interaction, identification of pain, depression and neurological disorders. This dissertation investigates the methods for facial expression analysis of people suffering from traumatic brain injury (TBI) and develops the system based on artificial emotional intelligence (AEI) for practical applications.

This dissertation focuses on the extraction of emotional signals from the patients suffering with TBI using computer vision techniques and the use of a social robot "Pepper robot" to assist in the rehabilitation. The work is organized into three themes: first, multimodal data collection from patients suffering from brain injury; second, extraction and recognition of facial emotions. Finally, the dissertation illustrates how extracted emotional signals can be applied in the effective human-robot interaction with the purpose of rehabilitation and social interaction.

Emotional signal extraction from the patients with brain injury is complex procedure due to unique and diverse psychological, physiological, and behavioral issues such as non-cooperation, face and body muscle paralysis, upper or lower limb disabilities, cognitive, motor, and hearing capabilities inhibition. It is necessary to interpret subtle changes in the emotional signals of people with brain injury for successful communication and implementation of affect-based strategies.

For data collection from subjects with brain injury, three different camera sensors, RGB, thermal, and depth, are used. New methods are introduced to gather good quality data for facial emotional recognition (FER). The thesis also presents a face quality assessment method to ensure a high-quality database in the face-log system.

In Facial Emotional Recognition, this dissertation has three main contributions: (i) development of state-of-the-art deep learning architecture for the

extraction and analysis of emotional signals exploiting visual and temporal networks, (ii) exploration of different techniques for fusing facial features from various visual modalities to improve predictive knowledge for the final model, and (iii) implementation of deep transfer learning techniques to overcome the challenges associated with database acquired from the subjects.

Within the human-robot interaction, the Pepper robot has been introduced, equipped with a deep-trained model for emotion recognition. The study emphasizes the real therapeutic value for stroke rehabilitation supported with tools to provide assessment and feedback in the neuro centers.

This dissertation includes peer-reviewed four conference publications, one book chapter and a journal publication. In the dissemination activities, a peer-reviewed conference paper has been presented targeted towards the design of customized tools for people with disabilities to enhance their communication for facilitating cognitive and physically impaired people.

Resumé

Mennesker er i stand til at producere et stort antal ansigtsudtryk (FE) ved at bruge deres ansigtsmuskler. Genkendelse af ansigtsudtryk (FER) er et aktivt forskningsområde, hvor menneskelige følelser bestemmes ved at klassificere ansigtsmuskelbevægelser. Automatisk genkendelse af ansigtsudtryk er af stor betydning for en lang række applikationer såsom biometrisk sikkerhed og overvågning, interaktion mellem menneske-maskine eller menneske-robot, identifikation af smerte, depression og neurologiske lidelser. Denne afhandling undersøger metoder som bruges til at analysere ansigtsudtryk af mennesker der lider af traumatisk hjerneskade (TBI) og udvikler et praktisk anvendeligt system baseret på kunstig følelsesmæssig intelligens (AEI).

Denne afhandling fokuserer på at analysere ansigtsudtryk fra patienter der lider af TBI ved hjælp af "Computer Vision" samt ved at bruge en social robot "Pepper-robot" til at hjælpe med rehabilitering. Arbejdet er organiseret i tre temaer: først, multi-modal dataindsamling fra patienter, der lider af hjerneskade; dernæst, genkendelse af ansigtsfølelser. Til sidst illustrerer afhandlingen, hvordan følelsesmæssige signaler som ansigtsudtryk kan anvendes i en effektiv menneske-robot-interaktion med formål om rehabilitering og social interaktion.

Genkendelse af ansigtsudtryk hos patienter med hjerneskade er en kompleks proces på grund af unikke og forskelligartede psykologiske, fysiologiske og adfærdsmæssige problemer såsom manglende samarbejde, lammelse af ansigts- og kropsmuskler, bevægelsesbesvær af øvre eller nedre lemmer, samt hæmmet hørelse eller reducerede kognitive og motoriske evner. Det er nødvendigt at fortolke subtile ændringer i de følelsesmæssige signaler fra mennesker med hjerneskade for vellykket kommunikation og implementering af affektbaserede strategier.

Til dataindsamling fra personer med hjerneskade anvendes tre forskellige kamerasensorer; RGB, termisk og dybde. Nye metoder introduceres til at indsamle data af god kvalitet til genkendelse af følelsesmæssige udtryk i ansigtet (FER). Afhandlingen præsenterer også en ansigtskvalitetsvurderingsmetode for at sikre en database af høj kvalitet i log-systemet af ansigter.

I FER har denne afhandling tre hovedbidrag: (i) udvikling af "Deep

Learning"-baserede modeller til ansigtsgenkendelse baseret på visuel og tidsafhængige data, (ii) udforskning af forskellige metoder til at kombinere ansigtsinformation ved brug af forskellige visuelle modaliteter med formålet at forbedre den beregnede forudsigelse i den endelige model, og (iii) implementering af "Deep Transfer Learning" for at overvinde de udfordringer, der er forbundet med den indsamlede database. Inden for menneske-robot interaktion anvendes en "Pepper-robot" udstyret med et neural netværk til ansigtsgenkendelse. Undersøgelsen understreger den reelle terapeutiske værdi for rehabilitering af slagtilfælde understøttet med værktøjer til at give vurdering og feedback i neurocentrene.

Denne afhandling inkluderer peer-reviewed fire konferencepublikationer, et bogkapitel og en tidsskriftpublikation. I formidlingsaktiviteterne er der præsenteret et fagfællebedømt konferenceoplæg, der er målrettet mod design af tilpassede værktøjer til handicappede for at forbedre deres kommunikation til at lette kognitive og fysisk handicappede.

Det sidste emne præsenterer en fagfællebedømt videnskabelig artikel, og er målrettet mod at designe og tilpasse værktøjer som kan forbedre kommunikation for kognitive og fysisk handicappede.

Thesis details

Thesis Title:	Facial Emotion Recognition for Citizens with Traumatic Brain Injury for Therapeutic Robot Interaction
PhD Supervisors:	Dr. Matthias Rehm Aalborg University Professor Kamal Nasrollahi Aalborg University
Submission date:	July 29, 2021

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published collection of papers listed below. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The main body of the thesis consist of following papers.

Emotion Recognition Analysis

[A] Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., & Moeslund, T. B. (2018, January). Facial Expression Recognition for Traumatic Brain Injured Patients. In VISIGRAPP (4: VISAPP) (pp. 522-530).

[B] Ilyas, C. M. A., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2018, October). Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 2291-2295).

[C] Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., & Moeslund, T. B. (2018, January). Effective Facial Expression Recognition Through Multimodal Imaging for Traumatic Brain Injured Patient's Rehabilitation. In International Joint Conference on Computer Vision, Imaging and Computer Graphics (pp. 369-389).

[D] Ilyas, C. M. A., Nunes, Rita., Rehm, M., Nasrollahi, K., & Moeslund, T. B. (2021, January). Deep Emotion Recognition through Upper Body Movements and Facial Expression. In 16th International Joint Conference on Computer Vision, Theory and Applications (VISAPP 2021).

[E] Ilyas, C. M. A., Rehm, M., Nasrollahi, K.,Yeganeh Madadi, Moeslund, T. B. & Vahid Seydi. Deep Transfer Learning in Human-Robot Interaction for Cognitive and Physical Rehabilitation Purposes. In journal of Pattern Analysis and Applications (PAA) with special issue on Computer Vision and Machine Learning for Healthcare Applications (PAA 2021).

Human-Robot Interaction

[E] Ilyas, C. M. A., Rehm, M., Nasrollahi, K.,Yeganeh Madadi, Moeslund, T. B. & Vahid Seydi. Deep Transfer Learning in Human-Robot Interaction for Cognitive and Physical Rehabilitation Purposes. In journal of Pattern Analysis and Applications (PAA) with special issue on Computer Vision and Machine Learning for Healthcare Applications (PAA 2021).

[F] Ilyas, C. M. A., Schmuck, V., Haque, M. A., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2019, October). Teaching Pepper Robot to Recognize Emotions of Traumatic Brain Injured Patients Using Deep Neural Networks. In 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1-7).

Dissemination Activities

[G] Ilyas C.M.A., Rodil K., & Rehm M. (2020) Developing a User-Centred Communication Pad for Cognitive and Physical Impaired People. In: Brooks A., Brooks E. (eds) Interactivity, Game Creation, Design, Learning, and Innovation. ArtsIT 2019, DLI 2019. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 328.

[H] Ilyas, C. M. A., Schmuck, V., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2018, December). 7th Aalborg U Robotics Workshop

Preface

This thesis is submitted as a collection of papers in fulfillment of a Ph.D. study at the Section of Media Technology, Aalborg University, Denmark. The work consists of research on three main topics: multimodal facial database development, deep learning architecture development for facial emotion recognition, and the intervention of a social robot "the Pepper robot" for rehabilitation of patients suffering from traumatic brain injury. Therefore, this dissertation is structured in four sections. Part one is the introduction and overview of the state-of-art in the three main topics as well as the contributions which have been made to them during this work. This is followed by a part containing selected papers published during this Ph.D. The last part offers the dissemination activities.

This project has been carried out from 2017-2020, with the collaboration of Osterskoven Neurocenter, Hobro, Denmark; Visual Analysis and Perception (VAP) Lab and Human-Robot Interaction (HRI) Lab at Aalborg University. In truth, I could not have achieved this level of success without strong mentors, encouraging friends, and my loving family. I thus first and foremost want to thank my mentor and supervisor, Prof. Dr. Matthias Rehm, for his excellent guidance throughout my Ph.D. He has given me all the freedom and encouragement needed to pursue my research interests while ensuring that I stayed on track. I am also grateful to my Co-supervisor Prof. Kamal Nasrollahi and head of Visual Analysis of People (VAP) Lab, Prof Thomas B. Moeslund, who always showed a profound interest in my research while providing enthusiastic guidance. An enormous thanks to all my current and previous colleagues in VAP and HRI , particularly Muhammad Ahsan ul Haque and Anne Jhuler Hansen, Ivan, and Kasper Hald.

Chaudhary Muhammad Aqdus Ilyas
Aalborg University,

Acknowledgements

Pursuing this Ph.D. has indeed been a tremendous life-changing experience for me that would not have been achieved without the support and guidance that I received from many people. First and foremost, I would like to express my sincere appreciation to my Ph.D. supervisors Dr. Matthias Rehm and Prof. Kamal Nasrollahi, for their excellent supervision and encouragement, which not only fostered my technical skills, but also edified my personality traits. During all these years, their profound wisdom, exceptional direction, outstanding passion for research, persistence and faith have been great source of inspiration to me.

My sincere gratitude also goes to Prof. Thomas Moeslund, head of Visual Analysis and Percepion (VAP) Lab, Aalborg University, who provided me opportunity to joining their team. I also thank Prof. Mohammed Bennamoun and Prof. Ajmal Mian for encouraging me to pursue my research abroad at University of Western Australia (UWA).

I am also grateful to the memebers of the doctoral committee, Associate Prof. David Meredith, Prof. Britta Wrede, and Prof. Drik Heylen for their in-depth study of this thesis, as well as for their informative questions and comments during my defense.

I am profoundly thankful to my colleagues and friends at the Aalborg University for all the memorable moments that we created together in course of my Ph.D. I also had several great moments with Pakistani friends during my stay in Aalborg.

Words cannot explain how grateful I am to my mother Tanveer, my father Muhammad Ilyas, and my brother and sisters, for all the unconditional love and sacrifices they have made for me.

Most significantly, I wish to thank my beloved wife Hijab and our daughter Aaira, from the core of my heart for all their continuing unconditional love, motivation and commitment. It is certainly with their absolutely unparalleled love that my PhD has been successfully completed.

Chaudhary Muhammad Aqdus Ilyas
Aalborg University, July 29, 2021

Contents

Part I

Overview of the work

Chapter 1

Introduction

Neurological disorders affect millions of people and contribute to worldwide death and disability. In particular, Traumatic Brain Injury (TBI) is one of the most common severe brain disorders and influences approximately 69 million people worldwide each year [5]. TBI is referred as an injury to the brain due to external sources and individuals sustaining a traumatic brain injury exhibit cognitive, motor, and behavioral challenges [39]. These imprecise functioning of the brain is not caused by neuro-degenerative or congenital/neuro-developmental conditions which means they were healthy before the stroke. TBI ranges from mild to moderate to severe, categorized by alteration of mental states, loss of consciousness, memory loss or lack of loco-motor coordination [15]. TBI severity is determined by the Glasgow Comma Scale (CGS) by measuring response to stimuli through eye opening, communication and motor activity [40]. Each year 5.48 million people suffer from the severe TBI that results in long-term disability [5]. TBI impose economic strain on the individual and community and can demonstrate devastating influence to ability of the affected persons to return to their families and perform social and occupational responsibilities [39]. Therefore, the neurological centers or the care units take care of these patients, where clinicians spend much time and resources rehabilitating them. However, apathy and decreased self-awareness are familiar characteristics associated with post traumatic brain injury, and have significantly negative influence on the process of rehabilitation. The study presents the technological aid in facilitating the staff members in retraining the residents by the Pepper robot's intervention and assessing the facial emotions.

TBI impacts negatively on the social, physical, and physiological interactions of the patients [37]. Therefore, social signal extraction from the residents with brain injury is extremely challenging. They have *"restricted or limited muscle movements, with reduced facial expressions and non-cooperative behavior, impaired reasoning, and inappropriate responses"* [17] with severe challenges regarding social communication and daily life activities. Almost 6 million individuals require extensive rehabilitation initiatives for their recovery, and several care attendants to look after them [39]. The primary issue in our studies is social signal extraction from people with brain injury for meaningful interaction and communication. The precision, speed, and accuracy of such social signals contribute vitally to the proper understanding of mental conditions and social

communication. This is not an easy task for care personnel and requires extended interactions with and intimate knowledge of the individuals. In addition, it is also a challenge to provide a technical solution.

This project presents the solution in two phases. The first phase is about the extraction of emotional or social signals from neurological impaired residents using computer vision techniques. The second phase integrates the emotion extraction system through the social robot for rehabilitation activities and monitors and assesses the emotions and retrained activities.

1 Emotion Recognition

Human emotions plays a vital role in human-human and human-machine interaction. Emotions represent the instantaneous mental states which vary according to human behavior and communication. Researchers are putting great emphasis on automatic recognition of human emotions as it is one of the essential parameter for natural human-machine interaction. In case of human-machine interaction, interaction would be impaired if machines are not able to recognize or understand the human emotions. This also applies to human-human interaction if other communication partners fail to understand these body expressions.

"Human recognize and demonstrate emotions through multi-modalities such as through facial expressions [2, 26, 27, 32, 35], body movements [4, 10, 31], speech recognition [2, 27, 35, 41] and physiological signals [1, 24, 28, 33, 38]" [21]. Facial expressions are primary sources of communication *"for human emotions, as approximately 55 percent of human communication happens through facial expressions [6, 34]"* [17]. Facial expressions exhibit the clues of mindset, intention and mood [6–8]. In addition, when there is a mismatch between facial expressions and other conversation medium, researchers weigh facial expressions more than other non-verbal communication channels in decoding emotions [3, 9]. Therefore, fast and accurate extraction of facial expressions can be beneficial in understanding social signals. Extraction of facial expressions and their interpretation in social signals is particularly challenging for patients with TBI. This is because facial shape and structure of muscles of these patients are different from healthy people, and therefore existing technologies, like those in [25, 30, 36] for extraction of facial expression from healthy people are challenged when applied to these patient [11–13, 23, 29]. This difference makes developing a system that works for each patient even more challenging.

Besides facial expressions, some other signals (including physiological, psychological, or physio-psychological signals), like a heartbeat, respiratory, fatigue, pain, and stress, can also be extracted from facial images and videos [1, 11–14, 24, 28, 33, 38]. These signals, similar to facial expressions, have a direct relationship with our daily activities and mood. However, repeatedly existing technologies for extracting these signals from facial images are not only targeting healthy people but are also limited to the lab environment [11–14, 23, 25, 29, 30, 36]. Therefore, the direct application of the existing technologies to TBI patients is not possible. Furthermore, to use these techniques properly for communication with patients, we need a reliable real-time system to extract these signals.

2 Human-Robot Interaction

Along with a robust facial expression system, a robotic framework can constitute a crucial set of assistive technology that facilitates and triggers/stimulates information for staff members and patients with brain injuries. Although the robotic interaction depends upon the facial expression evaluation, the understanding of the context is significant. The thesis investigates the video data collection, training a deep learning architecture, and interpreting social signals for the robotic interaction with people with TBI in the neurocenter.

The Ph.D. thesis aims to develop a precise and efficient system for the extraction of facial expressions, physiological and psychological signals, and interpreting them as social signals in the context of robotic interaction.
In human-robot interaction, the recognition of facial expressions is vital. New HRI systems apply various decision-making algorithms for the facial-emotion-perception unit, employing both static and temporal emotion states of the user. Computer vision is a powerful tool for real-time extraction of emotional signals; therefore, we have used the vision sensors to diagnose and monitor residents of a neurocenter. Thus, we aim to embed our system to extract facial, physiological, and psychological signals in a robot that can communicate with the residents. Therefore, this project is significant from a scientific point of view as it is the combination of numerous investigation grounds like computer vision, robotics and kinematics, machine learning, deep learning, psychology, and physiology. Finally, the Ph.D. study results include the development of a decision-based system using images and videos in real-time that can interpret the extracted signals for the understanding of the patient. The application in a rehabilitation context is exemplified by integrating it into a socially assistive robotic system and a pipeline of image and video analysis.

This research study is an essential step towards developing computer vision techniques for monitoring and diagnosing patients with brain injury by analyzing facial emotions. This exploration presents important auxiliary unbiased data to evaluate brain injury. The computer-aided system will empower us to determine consistent rehabilitation features that might help reduce the burden on the clinical staff members. The robotic system aids the clinical staff in decision making and saves time in providing clinical assessment. The robotic framework also contributes to the motivation of the residents for retraining activities and enhancing social communication.

3 Research Aims

Emotion recognition through facial expressions depends upon the facial muscle movement. However, patients suffering from brain injury exhibit varied muscle movements considering the impact of a neurological disorder such as paralysis, Bell's Palsy, ataxia, aphasia, and sensory-motor impairments. In this thesis, the focus is to address the challenges for facial expression recognition, for the emotional understanding of patients with neurological disorders and development of an affective-autonomous system for monitoring and assessing the patients. and facilitating the clinicians to enhance therapeutic human-robot interaction. This thesis focus on the vision-based

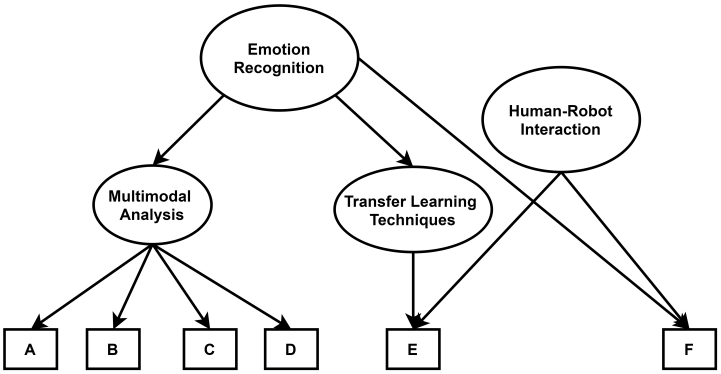


Fig. 1.1: Overview of articles in this thesis exploring emotion recognition and human-robot interaction. A box corresponds to a paper and a letter to the corresponding position in the appendix.

techniques, considering the non-intrusive visual sensors in a realistic environment. Emotion analysis of patients acquiring brain injury involves the static and dynamic characteristics captured through multi-modal visual sensors in a specialized neuro-care center registered during physical, cognitive, and social rehabilitation activities. In the context mentioned above, the dissertation focuses on five main objectives:

1) **Data Acquisition:** Identifying methods to acquire high quality data from the residents of the neurocenter to recognize emotional and physiological signals during activities of daily living (ADL). Studying the collected data with an expert skilled in communicating with these residents to annotate the above-collected video data with social signals (ground truth).
2) **Social Signal Extraction:** Depending on the number of different social signals identified in the above-captured data (with the help of the skilled communication expert), developing new and/or adapting of existing computer vision methods for real-time measurement of facial expression, physiological and psychological parameters for the individuals suffering from brain injury.
3) **Face Quality Assessment:** Studying the importance of different imaging conditions on the performance of the different developed algorithms. These paradigms could be, for example, facial pose, lightning environment/conditions, presence of facial elements on the face, and others. We aim to provide a high quality of the data to optimize the performance and lower the computation cost.
4) **Facial Expression Recognition System:** Having developed the systems mentioned in 3 for the signals identified in 2, develop a decision-based system that can interpret the collected signals to understand the mental states and behavior of the residents.
5) **Robotic Application of Facial Expression Recognition:** Embedding the developed systems in the robotics context and testing them in the field to analyze the gestures and reactions of residents before, during, and after the rehabilitation activities. Facilitate the staff members to make an informed decision based on monitoring and assessment by the socially-aware robotic-interaction systems.

In addition to the aforementioned five major challenges, this dissertation also presents the development of supporting tool for neurological impaired persons to have a better communication. Besides the facilitation, the study demonstrate the ways to deal with data scarcity using transfer learning techniques.

3.1 Thesis Contributions

In this dissertation a number of original contributions are made in the fields of computer vision and social robotics. The contributions are presented in Part II and Part III of the thesis.

Part II- Emotion Recognition of Traumatic Brain-Injured Patients

In the context of emotion recognition of people suffering from TBI, this thesis proposes the following main contributions: i) a new database for facial emotion recognition in three different modalities RGB (visible), thermal and Depth; ii) Extraction of social signals with the help of experts and care-providers in three specified scenarios such as physiotherapy, cognitive training and social rehabilitation; iii) High quality data acquisition by removing artifacts due to illumination, head and pose variation; iv) implementation of deep neural networks to extract facial features for facial expression analysis and their classification in to various emotional cues exploiting both spatial and temporal characteristics. A summary of the contributions are described as follows:

1. **Traumatic Brain-Injured Patients Facial Database** A new facial database has been presented, called TBI-database, including data from 11 subjects, captured with three different sensors. This database is collected in more than 30 sessions in collaborating neurocenters with each subject performing cognitive, physio and social communication activities. This database could not be published due to privacy issues of the patients, but this database has contributed in all publications made by the author.

2. **Facial Expression Recognition Based on a VGG + LSTM**
   A facial expression recognition solution based on the linear combination of VGG + Long Short Term Memory (LSTM), has been presented, exploiting the spatio-temporal information in a facial image of people suffering from TBI. This idea uses deep CNN representation to take advantage of complementary information of facial features identification, followed by the expression recognition. The VGG-Face descriptor, trained over 2.6 million face images, is built on VGG-16 network discarding the last fully connected layer in the architecture, to acquire 4096 feature vectors. This research work contributed to the following publication [18].

   - Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., & Moeslund, T. B. (2018,January). Facial Expression Recognition for Traumatic Brain Injured Patients. In VISI-GRAPP (4: VISAPP) (pp. 522-530).

3. **Facial Emotional Recognition Based on a VGG + LSTM with Multimodal Fusion**
   A facial expression recognition solution based on the linear combination of VGG + Long Short Term Memory (LSTM), has been presented, exploiting the

spatio-temporal information in the visible and thermal facial images of people suffering from TBI. The two different methods to fuse features are used to exploit the CNN - LSTM representations. On contrary to short-term 6 basic expressions, cognitive states and moods are interpreted, thus offering more socially-specific cues. This research work lead to the following publications [16, 19]

- Ilyas, C. M. A., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2018, October).Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 2291-2295).
- Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., & Moeslund, T. B. (2018, January). Effective Facial Expression Recognition Through Multimodal Imaging for Traumatic Brain Injured Patient's Rehabilitation. In International Joint Conference on Computer Vision, Imaging and Computer Graphics (pp. 369-389).

4. **Emotional Recognition Based on a VGG + LSTM through Facial Expressions and Upper Body Movements**
   This work proposes a new solution for emotion recognition through the combination of facial features and upper body features, computed in frame based and sequence based manner with different fusion techniques. The performance of proposed combined-features solution competes with state-of-the-art methods. This research work lead to the following publications [20]

   - Ilyas, C. M. A., Nunes, R., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2020, December). Deep Emotion Recognition through Upper Body Movements and Facial Expression. In 16th International Conference on Computer Vision Theory and Application (p. 229).

Part III- Human Robot Interaction for Social Rehabilitation of Traumatic Brain Injured Patients

In the context of human robot interaction and social rehabilitation of people suffering with TBI, this thesis proposes two main contributions: i) Implementation of deep-trained model in the Pepper robot application for the emotion recognition of people suffering with TBI; ii) Use of Deep Transfer Learning (TL) techniques to build the emotion recognition model for social robot interaction and rehabilitation; iii) Intervention of Pepper robot with aim to enhance social interaction and rehabilitation process. A summary of these contributions are presented as following

1. **Pepper Robot Intervention for Facial Emotion Recognition**

   To better understand the technological capabilities of the Pepper robot in the area of emotion recognition of the people suffering from TBI, this work proposes a comparative analysis of the deep-trained system for TBI database with built-in system of the Pepper robot to extract facial features and classify emotions and gesture synthesis accordingly. This research work lead to the publication of the following publication [22].

- Ilyas, C. M. A., Schmuck, V., Haque, M. A., Nasrollahi, K., Rehm, M., & Moeslund, T. B. (2019, October). Teaching Pepper Robot to Recognize Emotions ofTraumatic Brain Injured Patients Using Deep Neural Networks. In 2019 28th IEEE In-ternational Conference on Robot and Human Interactive Communication (RO-MAN)(pp. 1-7).

2. **Transfer Learning Techniques for Facial Emotion Recognition and Pepper Robot Gestures Synthesis for Rehabilitation Purposes**

   Following the proposed solution in 1. a comprehensive investigation on eight public datasets, recorded in-the-lab setting and in-the-wild settings, is made and applied the deep transfer learning techniques to fine-tune the TBI datasets for emotional recognition. Following the pilot studies, a Wizard-of-Oz (WoZ) functionality is introduced to facilitate the care-providers to achieve the rehabilitation targets with intellectual cognitive decision based on contextual information. This research work lead to the publication of the paper E.

   - Ilyas, C. M. A., Rehm, M., Nasrollahi, K.,Yeganeh Madadi, Moeslund, T. B. & Vahid Seydi. Deep Transfer Learning in Human-Robot Interaction for Cognitive and Physical Rehabilitation Purposes. In journal of Pattern Analysis and Applications (PAA) with special issue on Computer Vision and Machine Learning for Healthcare Applications (PAA 2021).

4 Thesis Structure

This chapter featuring a brief description of the context, motivation, research aims and contributions, followed by an appendix of three parts, each comprising a collection of papers within a specific field.

Part II presents the work conducted in the thesis on Emotion Recognition. As demonstrated in Figure 1.1, this involves four papers on multi-modal emotion analysis of neurological impaired people and one paper on transfer learning techniques for facial expression recognition.
Part III involves the research on human-robot interaction, compromising two research articles. Part IV presents the dissemination activities, more explicitly developing a user-centred communication pad for the rehabilitation of the physical and cognitive impaired people, published in 8th EAI International Conference: ArtsIT, Interactivity & Game Creation; and identifying effective data collection techniques for robotics to identify the facial expressions of neurological impaired people, an abstract published in 7th Aalborg U Robotics Workshop.

References

[1] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2011.

[2] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert)." *Emotion*, vol. 9, no. 5, p. 691, 2009.

[3] P. Carrera-Levillain and J.-M. Fernandez-Dols, "Neutral faces in context: Their emotional meaning and their function," *Journal of Nonverbal Behavior*, vol. 18, no. 4, pp. 281–299, 1994.

[4] B. de Gelder, A. De Borst, and R. Watson, "The perception of emotion in body expressions," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 2, pp. 149–158, 2015.

[5] M. C. Dewan, A. Rattani, S. Gupta, R. E. Baticulon, Y.-C. Hung, M. Punchak, A. Agrawal, A. O. Adeleye, M. G. Shrime, A. M. Rubiano *et al.*, "Estimating the global incidence of traumatic brain injury," *Journal of neurosurgery*, vol. 130, no. 4, pp. 1080–1097, 2018.

[6] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[7] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[8] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[9] J.-M. Fernández-Dols, H. Wallbott, and F. Sanchez, "Emotion category accessibility and the decoding of emotion from facial expression and context," *Journal of Nonverbal Behavior*, vol. 15, no. 2, pp. 107–123, 1991.

[10] S. Fridenson-Hayo, S. Berggren, A. Lassalle, S. Tal, D. Pigat, S. Bölte, S. Baron-Cohen, and O. Golan, "Basic and complex emotion recognition in children with autism: cross-cultural findings," *Molecular Autism*, vol. 7, no. 1, p. 52, 2016.

[11] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial video-based detection of physical fatigue for maximal muscle activity," *IET Computer Vision*, vol. 10, no. 4, pp. 323–329, 2016.

[12] ——, "Heartbeat rate measurement from facial video," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 40–48, May 2016.

[13] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain?" in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, Feb 2017, pp. 1–8.

[14] ——, *Heartbeat Signal from Facial Video for Biometric Recognition*. Cham: Springer International Publishing, 2015, pp. 165–174.

[15] J. Head, "Definition of mild traumatic brain injury," *J Head Trauma Rehabil*, vol. 8, no. 3, pp. 86–87, 1993.

[16] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 369–389.

[17] ——, "Facial expression recognition for traumatic brain injured patients," in *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, INSTICC. SciTePress, 2018, pp. 522–530.

[18] ——, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[19] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[20] C. M. A. Ilyas, R. Nunes, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Deep emotion recognition through upper body movements and facial expression," in *16th International Conference on Computer Vision Theory and Application*. SCITEPRESS Digital Library, 2020, p. 229.

[21] C. Ilyas, M. Rehm, K. Nasrollahi, Y. Madadi, T. Moeslund, and V. Seydi, "Deep transfer learning in human-robot interaction for cognitive and physical rehabilitation purposes," *Pattern Analysis and Applications*, 2021.

[22] C. M. A. Ilyas, V. Schmuck, M. A. Haque, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Teaching pepper robot to recognize emotions of traumatic brain injured patients using deep neural networks," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[23] R. Irani, K. Nasrollahi, A. Dhall, T. B. Moeslund, and T. Gedeon, "Thermal superpixels for bimodal stress recognition," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.

[24] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE, 2011, pp. 410–415.

[25] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, Feb 2014.

[26] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.

[27] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, and N. Boulanger-Lewandowski, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[28] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[29] J. Klonovs, M. A. Haque, V. Krueger, K. Nasrollahi, K. Andersen-Ranberg, T. B. Moeslund, and E. G. Spaich, *Monitoring Technology*. Cham: Springer International Publishing, 2016, pp. 49–84.

[30] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, Jan 2007.

[31] K. Lang, M. M. Dapelo, M. Khondoker, R. Morris, S. Surguladze, J. Treasure, and K. Tchanturia, "Exploring emotion recognition in adults and adolescents with anorexia nervosa using a body motion paradigm," *European Eating Disorders Review*, vol. 23, no. 4, pp. 262–268, 2015.

[32] L. Y. Mano, B. S. Faiçal, L. H. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, P. Geraldo Filho, G. T. Giancristofaro, G. Pessin, B. Krishnamachari *et al.*, "Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition," *Computer Communications*, vol. 89, pp. 178–190, 2016.

[33] A. Martínez-Rodrigo, R. Zangróniz, J. M. Pastor, J. M. Latorre, and A. Fernández-Caballero, "Emotion detection in ageing adults from physiological sensors," in *Ambient Intelligence-Software and Applications*. Springer, 2015, pp. 253–261.

[34] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 1.2, no. 4, pp. 53–56, 1968.

[35] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–1, 2017.

[36] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, April 2006.

[37] R. S. Parker, *Traumatic brain injury and neuropsychological impairment: Sensorimotor, cognitive, emotional, and adaptive problems of children and adults*. Springer Science & Business Media, 2012.

[38] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.

[39] M. Sherer and A. M. Sander, *Handbook on the neuropsychology of traumatic brain injury*. Springer, 2014, vol. 144.

[40] G. Teasdale and B. Jennett, "Assessment of coma and impaired consciousness: a practical scale," *The Lancet*, vol. 304, no. 7872, pp. 81–84, 1974.

[41] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

Chapter 2

Emotion Recognition of Traumatic Brain-Injured Patients

1 Background

Social signal extraction from TBI patients is very challenging in health industries, which requires long lasting work of resources, machines and products. It also plays a vital role in communication and interaction enhancement, as brain disease results in restricted or limited muscle movements so people suffering from TBI face serious difficulty at social communication and activities of daily livings (ADL). There are almost 6 million individuals suffering from TBI [1] [13], who require large number of care attendants to look after them. In modern era, the rehabilitation and training of people suffering from brain injuries, is focused through the development of unique, customized and adjustable robotic systems [19, 42, 47, 56]. The main aim is to aid people with disabilities in the activities of daily living (ADL) and enhancing social interaction, resulting in better quality of life. Moreover, these robotic system can lessen the burden of specialized care homes or neuro-center as the number of people with disabilities is increasing day by day. *"According to the International Brain Injury Association (IBIA), only in America approximately one million people suffer from traumatic brain injury (TBI) annually, whereas the same number of people suffer from TBI in all over Europe [1]. The American Center for Disease Control and Prevention estimates more than 3.7 million people are living with long term disability after TBI [13]."* [31] Furthermore, the average age has been expanded to 75 - 85 years in most developed countries particularly in Europe, thus the average percentage of elderly people with special needs has been significantly increased. It is believed that in 2050, Europe's population will consist of 60 plus aged people up to 40% of its total population [51]. According to the United States Census bureau, over 65 years aged people make up 14.9% of USA population that is equal to 47.8 million people in year 2015 and this population of

elderly people is projected to double in 2060 with approximately 98.2 million, with 19.7 million will be over 85 years of age [2]. This will eventually lead to more care homes, neuro-centers and more staff to look after elderly people. Therefore, there is an increased demand of special robots that perform tasks like carer, companion, health monitor, educator, trainer and so on.

The real challenge is how we can place robots in real time applications with special target groups like people with acquired brain injuries who are already suffering from cognitive, functional, sensory, psychological, intellectual and behavioral impairments. They can have agitation, confusion, loud verbalization as well as physical aggression [39]. In addition to that, for patients suffering from TBI, robots must be customized according to the nature of the disability, as some individuals suffer with paralysis, speech inhibition, reduced expressions, limited hand-eye coordination and etc. The main idea is designing autonomous or semi autonomous robots that aid in training cognitive and physical skills with minimal human supervision. In our thesis we have investigated a frame work for a robotic application that accurately identifies the facial expressions and their interpretations as social signals to be used in human-robot interaction scenarios. The first step in this process is database acquisition, essential for developing deep learning system to identify facial expressions.

2 Data Acquisition System

For the purpose of well-being of neurologically impaired people through emotional signals, we need to acquire the relevant data. Amygdala, one of the region of the brain generates emotional signals and their corresponding reactions in a number of ways such as: heart rate, blood pressure, skin conductance, pupil-dilation, brain activity and facial expressions [3, 8, 46].
Facial expression is one of the most effective way to understand emotions and non-verbal communications [15–17]. Therefore, to estimate the emotional cues with unobtrusive sensors, we decided to collect the facial data in three different modalities such as RGB, thermal and depth. Data is acquired with Microsoft Kinect (visible and depth imagery), Axis Q1922 (thermal imagery) and Axis RGB-Q16 (visible) sensors. The following sections will discuss the characteristics of the sensors used to acquire data and data acquisition platform.

2.1 Microsoft Kinect for Windows V2

Microsoft Kinect 2 for windows is a second-generation camera released by Microsoft in 2014. It contains an RGB camera, an infrared laser emitter, an infrared camera, and a multi-array microphone. According to the technical specifications, Kinect V2 has 70 °and 60 °horizontal and vertical Field of View (FOV) as compared to 57 °and 43 °horizontal and vertical FOV respectively for Kinect V1. Additional specifications of Kinect V2 are presented in the Table 2.1.

Microsoft Kinect 2 for windows SDK has a capability to identify speech, track skeleton, and many methods for obtaining color and depth information. However, we require only raw color and depth data in this context.

Table 2.1: Micrsoft kinect V2 technical specifications [48]

Features		Values
Visible	Resolution	1920 x 1080 px
	Frame Rate	30 FPS
	Horizontal FOV	84.1°
	Vertical FOV	53.8°
Depth	Resolution	512 x 424 px
	Frame Rate	30 FPS
	Horizontal FOV	70.6°
	Vertical FOV	60°
Camera Range		0.5 - 4.5m
Standard USB		3.0

2.2 Axis Q1922-E Thermal Camera

The Axis Q1922-E is a thermal camera with infrared range 8-14μ m, able to visualize the objects with temperature ranges from -20 to 100°. The camera utilizes an uncooled micro bolometer to acquire the radiations [5]. Technical specification of Axis Q1922 is presented in the Table 2.2.

Table 2.2: Axis Q1922-E Thermal Camera technical specifications. *Frame rate is applicable to selected countries such EU, UK and USA. Adopted from [7] and [5]

Features	Values
Resolution	640 x 480 px
Frame Rate	30 FPS *
Spectral Range	8 - 14 μm
Lens	19mm

2.3 Axis RGB-Q16

The Axis Q1615 camera delivers high quality images with unique "Lightfinder technology" that enables to visualize the objects even in the low light conditions. It posses Electronic motion stabilization that increases the video quality when the camera sensor experience vibration. Technical specification of Axis Q1615 is presented in the Table 2.3.

2.4 Logitech C920

We have used a HD webcam camera Logitech C920 that is equipped with automatic light correction that fine-tune lighting conditions to provide bright and contrasting

Table 2.3: Axis Q1615 Camera technical specifications. [6]

Features	Values
Resolution	1920 x 1080 px
Frame Rate	25/30 FPS
Field of View	90°x 40°
Lens	2.8-8 mm

images in less illumination conditions. Technical specification of Logitech C920 is presented in the Table 2.4

Table 2.4: Logitech C920 technical specifications. [43]

Features	Values
Resolution	1920 x 1080 px
Frame Rate	30 FPS
Focus	Autofocus
Standard USB	2.0

2.5 Sensor Registration

For the multimodal analysis, it is essential to register the various sensors accurately. Thermal and visual modalities can be fused by pixel-level, feature-level and decision-level fusions. Researchers [7, 37, 65] have studied the methods for registration of modalities based on 3D scene information, registration methods, fusion methods and necessary assumptions for sensor registration. Following [37], we have used calibrated chess-board blob-homography technique. This approach assumes that foreground objects in a thermal image are hotter than background objects, eliminating the pixel-wise correspondence of all objects in the picture [7]. According to [37], blob homography performs accurately in a range from 95-99% in image rectification systems. For the accurate pixel level registration, following the notion of [7], we have used the depth information provided by Microsoft Kinect camera. However, to ensure that objects in both modalities moves synchronously, we have timely synchronized the both modalities.

2.6 Data Fusion

After the sensor registration and images alignment, we tried to fuse the image modalities in the various combinations such as early (feature-level) fusion, late (decision-level) fusion like [62] and bilinear pooling (point-wise multiplication a type of feature-level fusion). Early fusion use the features correlations from time-synchronous modalities and late fusion can used for asynchronous multimodal data thus providing flexibility to train on larger amount of available datasets. In our research dissertation we

Fig. 2.1: Hardware platform with Axis-RGB, Logitech C920, Axis Q1922-E and Microsoft Kinect V2 for the data collection

have performed early fusion by concatenating the feature vectors from both modalities and applied 10-folds cross validations to evaluate various subsets of features. In late fusion, the outcome of multiple classifiers or regressors are combined to make a final estimation by using maximum of posterior probabilities incorporating Sum and Product rule [12, 23].

3 Data Collection

Data collection took place during different rehabilitation activities and depending on the subject's challenges. We have collected data from eleven subjects suffering from TBI, detailed nature of their disability is described in table E.1 [29, 31, 33]. Subjects suffering with TBI have mild to severe stroke, followed with paralysis, speech inhibition and emotional instability. Based upon their health conditions and neuro-psychological test results, psychologists and neuro-rehabilitation experts devised strategies for recovery [9] [59]. Furthermore, the subjects under observations face difficulty to express their feeling and emotions, and most often have behavioral challenges such as mood swings, low concentration and sometimes verbal or physical aggression [29, 31, 33]. All previous mentioned factors contribute to the complexity in facial data collection. Challenges associated with data collection and adopted strategies to acquire good quality of data have been briefly discussed in our research articles [29, 31, 33]. However, in this chapter, we will discuss in detail.

First of all, for the uniformity of data collection, we have selected three scenarios where subjects perform activities for recovery in social communication; Cognitive rehabilitation activities, and physiotherapy sessions. However, it is observed that in above-mentioned activities, data for facial analysis can not be collected with predefined strategies and there is a need of adjustment to cater the needs of data driven facial analysis systems. Therefore, considering the visual system we modified the rehabilitation strategies. Details of standard rehabilitation strategy and modified pro-

cedures have presented below.

3.1 Cognitive Rehabilitation Strategy

Cognitive activities aim to train the mental functionalities and evaluate the patient's ability to understand and interpret the information. For the optimal training, it is essential to have stable emotional conditions of the subjects under observation, otherwise rehabilitation process could be lengthy and miss the targets. Therefore, care providers follow *" a set of protocols comprised of repetitive activities with gradual increase in difficulty levels like Mini-Mental State Exam (MMSE)*[1] *and Montreal Cognitive Assessment (MoCA)*[2]*, to assess attention, memory [54], visuo-spatial perception [44], language and communication, function execution and learning ability of the patient [59]"* [32]. Care-providers complete the required tasks by number of ways, for instance, use of memory devices or memory logs, clocks or calendars illustration, alarms or reminders, book-reading and listening and design concept mapping. However, all these tasks are tailored to the baseline level of the subject's understanding with gradual increase in difficulty. Unfortunately, aforementioned techniques does not support data driven autonomous systems to access the cognitive performance due to variable facial pose, gestures, and less attention during the execution of the tasks. In order to address the challenges, we modified these methods by switching to digital interface devices like tablets with webcam enabled to capture the frontal faces. subjects are then asked to watch a favourite movie clip and then talk about that character, listen and sign lyrics of songs, matching pictures activities, and playing games in order to achieve Error Less (EL) learning. Considering sad and depressed emotions associated with TBI patients for the most of the time, games are designed with more incentives and winning probabilities to generate happy emotions, motivations and increased attention. After approval from the care-providers these modified strategies were implemented on 11 subjects that resulted in memory enrichment and enhanced attention, core ingredient for effective cognitive training. Details have been presented in the paper E.

3.2 Physical Rehabilitation Strategy

Physical morbidity is caused by TBI when sensory motor skills are impaired. Reduced muscle coordination, upper limb, lower limb, or full body paralysis may all result from a stroke, depending on the severity. Functional therapeutic techniques are proposed on a case-by-case basis, considering factors such as age, gender, ethnicity, disability level, and post-concussion symptoms [27]. Furthermore, while developing and executing aerobic, musculoskeletal and neuromuscular exercises; subject's stability, coordination and neuromuscular functionalities estimation are considered. Physiotherapists do pre-set operations such as aerobic exercise, treadmill, walk or gentle running independently or with a trainer, swimming, bicycling, bench presses, squats, and other exercises for recovery objectives [27]. However, in all these activities, facial data that is essential for data driven assessment systems, is rarely available due

[1] https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/undersoegelser-og-proever/skemaer/geriatri/mms-mini-mental-status/

[2] https://www.mocatest.org/

to excessive face and body movements. Therefore, with aim of maximum facial data collection, we modified the physical activities depending upon the disability nature of the patients. For instance, subjects with upper body paralysis are devised to ride stationary bicycle to record facial expressions while placing a camera parallel to the face. Similarly, patients who are suffering with lower limb paralysis and bounded to the wheel chair, physical activity is designed to move wheel chair back and forth within three meters for multiple sessions. Moreover, some of the subjects moved their arms and hands to certain limits while playing console games or card games. Similarly, staff members and physiotherapist also designed special activities like blowing air in the bottle, use of hand-press, organizing plates on the table and other related activities with aim to capture more useful data and increase the interests of the participants.

3.3 Social Rehabilitation Strategy

People suffering with neurological disorders experience challenges in social communication social interaction, and social integration. Therefore, care providers adopt unique and customized strategies based on the individual mental functional capabilities and behavioral disorders. Conventional methods such as reading books or storey telling, sharing personal experiences or daily routines could not produce effective results at the neuro-center due to lack of interest, concentration and poor story telling skills. The aforementioned strategies were adopted to more interesting and engaging activities such playing cards games and video-console games according to the interest of the participating groups. It is observed that all the participants enjoyed and exhibited more interest in the modified activities and games. Some of the games such as Medal of Honor Airborne (MOHA)[3], Need for Speed[4], where coordination and group understanding is required, participants struggle to get good score, but resulted in more communication and social interaction. Details of these modified approaches and game designs have been presented in the appendix E.

4 TBI-Database

In the following section, we will provide an overview of the database collected from the people suffering from TBI. Data is collected in multiple phases during 91 sessions, as illustrated in table E.1 with RGB, thermal and depth sensors. We have recorded 1723 video events, each of maximum 5 seconds in length, resulting in approximately 250,000 image frames. However, it is observed that data collected with modified strategies was less erroneous when applied pre-processing techniques to log on to the face log system. Face Quality Assessment (FQA) is applied to the entire database to ensure high quality of data is fed into the deep neural networks.

Face Quality Assessment Method

To ensure high quality of the data to avoid computational cost of the erroneous data, we have employed the Face Quality Assessment (FQA) method [24, 29–31, 34]. FQA

[3]https://www.ea.com/games/medal-of-honor
[4]https://www.ea.com/games/need-for-speed

checks the low quality of images due to large pose variation, occlusion, intensity, sharpness and insufficient resolution and discards the images by comparing it with reference image before logging into facial log system. Details of FQA method has been presented in the [31].

Table 2.5: Subjects in database along with challenges due to TBI, number of sessions and activities [E]

Subjects	No. of Sessions	Activities			Challenges			Prominent Features
		Cognitive	Physio	Social	Body Paralysis	Speech Inhibition	Facial Paralysis	
A	12	4	4	4	Complete	Yes	Partial	High Anger
B	10	4	3	3	Left Side	No	No	High Arousal
C	10	4	3	3	Lower Body	No	No	Excessive Head Movement
D	9	3	3	3	Partial	No	Partial	Emotionally Unstable
E	9	2	4	3	No	Yes	Partial	Emotionally Unstable
F	7	2	3	2	Partial	No	No	High Arousal
G	6	2	2	2	Lower Body	No	No	Excessive Upper Body Movement
H	7	2	3	2	No	No	Partial	Low Arousal
I	6	2	2	2	Yes	Yes	Partial	Low Arousal
J	8	2	3	3	No	No	No	Verbal and Physical Aggression
K	7	3	3	1	Partial	Yes	No	Emotionally Unstable

5 Deep Learning based Emotion Recognition: State-of-the-Art

Emotions play a critical role in human communication, social interaction, cognitive judgement and human robot interaction. Human emotions can be described through the facial muscle movements (Facial Expressions), Electroencephalography (EEG), Electrocardiography (ECG), Magnetic resonance imaging (MRI) and skin conductance [8]. The simplest and the most wide spread method to identify emotions is through recognition of 6 basic facial expressions (Anger, Disgust, Fear, Happy, Sad and Surprise) proposed by Ekman et al [15] plus neutral expression, also called the discrete emotions. Facial emotional expressions can also be dimensionally categorized in terms of valence (measure of pleasantness) and arousal (measure of activation) [10, 22]. The combination of seven different expressions can be presented in valence-arousal space as illustrated in the figure 2.2. According to Mehrabian [45], more than 55% of human communication is performed through facial expressions. Automatic recognition of facial expressions plays a vital role in various applications in the fields of biometrics, forensics, health care, medical diagnosis, monitoring and surveillance [15, 28, 30, 40, 41]. Since past couple of decades, researchers have exerted a great effort to develop methods and systems for robust and accurate identification of emotions that go beyond the Ekman's basic facial expressions. Facial expressions present information to be interpreted as mood, cognitive states and emotional states. Therefore, the expressions are vital tools in the development of human-computer and human-robot interaction systems.

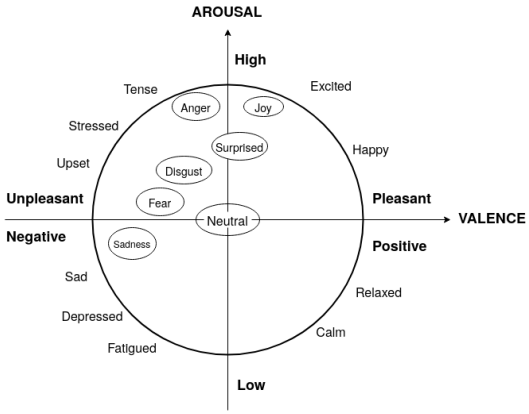


Fig. 2.2: Distribution of Ekman [15] emotions in valence-arousal space. Image is adopted from [10] and [22]

The current facial emotion recognition (FER) systems can be distinguished on the basis of facial feature extraction and classification techniques. Facial features can be extracted by using geometrical features, apparent features or hybrid (both geometric and apparent) features [30, 35, 50, 57]. *"Geometry based feature extraction methods use geometric shape and position of the facial parts like lips, nose, eyebrows and mouth, with temporal information such as the movement of facial features points from the previous frame to the current frame [21, 25]*" [30]. One of the advantages of using geometrical features is its tolerance against illumination variation and non-frontal head pose handling by measuring fiducial points from frontal images [4, 52]. Texture information of facial images is used in appearance based methods, whereas in hybrid methods, both geometric and apparent features are represented [52].

FER systems can also be categorized on the basis of features classification methods. For instance [21] have classified the facial features with the help of Support Vectors Machines (SVM), [60, 66] used Local Binary Patterns (LBP). Linear Discriminant Analysis (LDA) and Hidden Markov Models (HMM) are used in [61], Wavelet approaches in [49, 52, 63], Non-Negative Matrix Factorization (NMF) and Discriminant NMF in [14, 53]. Similarly some researchers have employed the combination of various methods to extract and classify facial features such as in [38], features are selected by the AdaBoost Algorithm and classified by Gabor filters. Researchers in [20] have extracted temporal facial information and categorized into emotions by combining AdaBoost and SVM classifier.

Systems build on Deep Learning (DL) approaches, that is based on Deep Convolutional Neural Networks (DCNN) have produced more accurate, fast and robust results, for facial features extraction and classification as compared to classical approaches [11, 18, 58, 64]. [36] have employed the Deep Belief Neural Networks for exploring facial expression and pain analysis. Similarly, [26] and [55] have determined facial expressions and pain assessment by linear combination of CNN and Long Short Term Memory Networks (LSTM) by incorporating spatio-temporal information and

achieved the state-of-the-art performances. [11] have improved this method by using super-resolved facial frames to train the CNN-LSTM network.

In our literature review, two main aims are acquired: identifying the transition and gap in the research area of emotion recognition from hand-crafted features methods to machine learning driven algorithms and recognising the assessment studies across facial expressions variations. According to the literature review conducted, we have also identified that most of the studies are performed on the healthy subjects databases with lesser or no facial artifacts in different environmental conditions. This research aims to identify the emotions of people suffering from TBI with extended physiological and physical challenges like facial and body paralysis. These emotions are inferred from various computer vision applications that understand the facial expressions, cognitive states such as interested or bored and mood analysis.

References

[1] Brain injury facts | international brain injury association-ibia. [Online]. Available: http://www.internationalbrain.org/brain-injury-facts/

[2] Facts for features: Older americans month: May 2017.

[3] R. Adolphs, "Neural systems for recognizing emotion," *Current opinion in neurobiology*, vol. 12, no. 2, pp. 169–177, 2002.

[4] R. N. Anwar Saeed, Ayoub Al-Hamadi and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in Human-Computer Interaction*, vol. 2014, pp. 1–13, 2014.

[5] *AXIS Q1922-E Thermal Network Camera*, Axis Communications, 02 2011, rev. 16.

[6] *AXIS Q1615-E Thermal Network Camera*, Axis Communications, 12 2014, rev. 4.5.

[7] C. Bahnsen, "Thermal-visible-depth image registration," *Unpublished Master Thesis, Aalborg University, Aalborg, Denmark*, 2013.

[8] M. Balconi, "Neuropsychology of facial expressions. the role of consciousness in processing emotional faces," *Neuropsychological Trends*, vol. 11, pp. 19–40, 2012.

[9] A. Barman, A. Chatterjee, and R. Bhide, "Cognitive impairment and rehabilitation strategies after traumatic brain injury," *Indian Journal of Psychological Medicine*, vol. 38, no. 3, pp. 172–181, May-Jun 2016.

[10] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition & Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[11] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*. Cham: Springer International Publishing, 2017, pp. 151–162.

[12] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 205–211.

[13] T. CA, B. JM, B. MJ, and X. L., "Traumatic brain injury–related emergency department visits, hospitalizations, and deaths — united states, 2007 and 2013," *Morbidity and Mortality Weekly Report (MMWR)*, vol. 66(No. SS-9), p. 1–16, 2017.

[14] G.-J. de Vries, S. Pauws, and M. Biehl, *Facial Expression Recognition Using Learning Vector Quantization*. Cham: Springer International Publishing, 2015, pp. 760–771.

[15] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J Pers Soc Psychol*, vol. 17, no. 2, pp. 124–129, Feb 1971.

[16] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[17] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[18] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ser. ICMR '15. ACM, 2015, pp. 643–650.

[19] R. Gassert and V. Dietz, "Rehabilitation robots for the treatment of sensorimotor deficits: a neurophysiological perspective," *Journal of neuroengineering and rehabilitation*, vol. 15, no. 1, pp. 1–15, 2018.

[20] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[21] D. Ghimire, J. Lee, Z.-N. Li, and S. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7921–7946, Mar 2017.

[22] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[23] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE international conference on systems, man and cybernetics*, vol. 4. IEEE, 2005, pp. 3437–3443.

[24] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Aug 2013, pp. 443–448.

[25] ——, "Constructing facial expression log from video sequences using face quality assessment," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, Jan 2014, pp. 517–525.

[26] ——, "Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain?" in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, Feb 2017, pp. 1–8.

[27] J. A. Hugentobler, M. Vegh, B. Janiszewski, and C. Q. Yates, "Physical therapy intervention strategies for patients with prolonged mild traumatic brain injury symptoms: A case series," *International Journal of Sports Physical Therapy*, vol. 10, no. 5, pp. 676–689, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4595921/

[28] M. P. Hyett, G. B. Parker, and A. Dhall, *The Utility of Facial Analysis Algorithms in Detecting Melancholia*. Cham: Springer International Publishing, 2016, pp. 359–375.

[29] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 369–389.

[30] ——, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[31] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[32] C. Ilyas, M. Rehm, K. Nasrollahi, Y. Madadi, T. Moeslund, and V. Seydi, "Deep transfer learning in human-robot interaction for cognitive and physical rehabilitation purposes," *Pattern Analysis and Applications*, 2021.

[33] C. M. A. Ilyas, V. Schmuck, M. A. Haque, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Teaching pepper robot to recognize emotions of traumatic brain injured patients using deep neural networks," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[34] R. Irani, K. Nasrollahi, A. Dhall, T. B. Moeslund, and T. Gedeon, "Thermal superpixels for bimodal stress recognition," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.

[35] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, Feb 2014.

[36] R. Kharghanian, A. Peiravi, and F. Moradi, "Pain detection from facial images using unsupervised feature learning approach," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 419–422.

[37] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 270–287, 2007.

[38] S. Lajevardi and Z. Hussain, "Novel higher-order local autocorrelation-like feature extraction methodology for facial expression recognition," *IET Image Processing*, vol. 4, pp. 114–119(5), April 2010.

[39] M. D. Lauterbach, P. L. Notarangelo, S. J. Nichols, K. S. Lane, and V. E. Koliatsos, "Diagnostic and treatment challenges in traumatic brain injury patients with severe neuropsychiatric symptoms: insights into psychiatric practice," *Neuropsychiatr Dis Treat*, vol. 11, pp. 1601–1607, 2015.

[40] F. Li, C. Zhao, Z. Xia, Y. Wang, X. Zhou, and G.-Z. Li, "Computer-assisted lip diagnosis on traditional chinese medicine using multi-class support vector machines," *BMC Complementary and Alternative Medicine*, vol. 12, no. 1, p. 127, Aug 2012.

[41] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, second edition ed. Springer London Dordrecht Heidelberg New York: Springer, 2011.

[42] A. C. Lo, P. D. Guarino, L. G. Richards, J. K. Haselkorn, G. F. Wittenberg, D. G. Federman, R. J. Ringer, T. H. Wagner, H. I. Krebs, B. T. Volpe *et al.*, "Robot-assisted therapy for long-term upper-limb impairment after stroke," *New England Journal of Medicine*, vol. 362, no. 19, pp. 1772–1783, 2010.

[43] *Logitech HD Webcam C920*, Logitech, 2011.

[44] K. McKenna, D. M. Cooke, J. Fleming, A. Jefferson, and S. Ogden, "The incidence of visual perceptual impairment in patients with severe traumatic brain injury," *Brain Injury*, vol. 20, no. 5, pp. 507–518, 2006. [Online]. Available: https://doi.org/10.1080/02699050600664368

[45] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 1.2, no. 4, pp. 53–56, 1968.

[46] R. M. Müri, "Cortical control of facial expression," *Journal of comparative neurology*, vol. 524, no. 8, pp. 1578–1585, 2016.

[47] K. Y. Nam, H. J. Kim, B. S. Kwon, J.-W. Park, H. J. Lee, and A. Yoo, "Robot-assisted gait training (lokomat) improves walking function and activity in people with spinal cord injury: a systematic review," *Journal of neuroengineering and rehabilitation*, vol. 14, no. 1, pp. 1–13, 2017.

[48] D. Pagliari and L. Pinto, "Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors," *Sensors*, vol. 15, no. 11, pp. 27 569–27 589, 2015.

[49] G. Palestra, A. Pettinicchio, M. Del Coco, P. Carcagnì, M. Leo, and C. Distante, *Improved Performance in Facial Expression Recognition Using 32 Geometric Features*. Cham: Springer International Publishing, 2015, pp. 518–528.

[50] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, April 2006.

[51] M. E. Pollack, "Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment." *AI Magazine*, vol. 26, no. 2, pp. 9–24, 2005. [Online]. Available: http://dblp.uni-trier.de/db/journals/aim/aim26.html#Pollack05

[52] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. N. Yanushkevich, "Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, p. 17, Sep 2012.

[53] A. Ravichander, S. Vijay, V. Ramaseshan, and S. Natarajan, *Automated Human Facial Expression Recognition Using Extreme Learning Machines*. Cham: Springer International Publishing, 2016, pp. 209–222.

[54] L. Rees, S. Marshall, C. Hartridge, D. Mackie, and M. W. F. T. E. Group, "Cognitive interventions post acquired brain injury," *Brain Injury*, vol. 21, no. 2, pp. 161–200, 2007. [Online]. Available: https://doi.org/10.1080/02699050701201813

[55] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE transactions on cybernetics*, 2017.

[56] B. Sheng, Y. Zhang, W. Meng, C. Deng, and S. Xie, "Bilateral robots for upper-limb stroke rehabilitation: State of the art and future prospects," *Medical engineering & physics*, vol. 38, no. 7, pp. 587–606, 2016.

[57] Y. I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, Feb 2001.

[58] D. Triantafyllidou and A. Tefas, "Face detection based on deep convolutional neural networks exploiting incremental facial part learning," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 3560–3565.

[59] T. Tsaousides and W. A. Gordon, "Cognitive rehabilitation following traumatic brain injury: assessment to treatment," *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 76, no. 2, pp. 173–181, 2009. [Online]. Available: http://dx.doi.org/10.1002/msj.20099

[60] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[61] M. Z. Uddin and M. M. Hassan, "A depth video-based facial expression recognition system using radon transform, generalized discriminant analysis, and hidden markov model," *Multimedia Tools and Applications*, vol. 74, no. 11, pp. 3675–3690, Jun 2015.

[62] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration-a statistical view," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334–341, 1999.

[63] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790 – 799, 2014, best of Automatic Face and Gesture Recognition 2013.

[64] H. Yoshihara, M. Seo, T. H. Ngo, N. Matsushiro, and Y. W. Chen, "Automatic feature point detection using deep convolutional networks for quantitative evaluation of facial paralysis," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Oct 2016, pp. 811–814.

[65] J. Zhao and S. C. Sen-ching, "Human segmentation by geometrically fusing visible-light and thermal imageries," *Multimedia Tools and Applications*, vol. 73, no. 1, pp. 61–89, 2014.

[66] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," *Sensors*, vol. 11, no. 10, pp. 9573–9588, 2011.

Chapter 3

Human Robot Interaction

1 Introduction

In recent years, due to the advancements in automation, robotics and artificial intelligence (AI), robots have exhibited a huge potential to support and assist humans in range of contexts such as homes [5, 15], workplaces [12, 36], education institutions [27], healthcare environment [4, 25, 38] and many more. The robotic technology has impacted humans positively in every sector of the life. In this dissertation, we offer an introduction of a assistive robot to support the healthcare workers, staff members, nurses, caregivers, trainers and therapists in a Danish neurocenter who are working with people suffering from neurological disorders. This study aims to *"investigate the technological innovations to assist residents and staff members with fulfillment of rehabilitation activities - including enhancing individual self-control and improvement of quality of life [18, 19, 32]"* [22].

2 Collaboration Partners

Aalborg University works closely with other institutions, industries and healthcare units to mitigate the challenges of deploying and appropriating interactive technology. Since 2015, AAU has worked closely together with staff and residesnts at Senhjerneskadecenter Fredrikshavn and Neurocenter Østerskoven in Hobro on co-designing and developing assisitive robot technology. Both of these centers are providing care and assistance to people who have acquired brain injuries. Due to the stroke, either naturally or by accident, results in impairments which leads to failure to perform activities of daily living (ADL). Depending upon the severity of the stroke, some of these residents have acquired life-long disability (in terms of cognitive, physical and communication), therefore, staff members or care givers have to look after them 24/7 and retrain them with set of skills necessary to perform ADL by repetitive execution of the tasks [22].

The Danish neurocenters focus on the people suffering from neurological disor-

ders to provide assistance to physical, social-communication and mental disabilities by following the set of protocols. For instance, neurocenters develops and adopts cognitive activities with an aim to *"improve the ability of residents to understand and interpret information to perform specific functions mentally"* [20] by assessing the mental stability and repetitive execution of tasks with gradual increase in difficulty. Similarly, physical retraining is planned by assessing the physical morbidity of the subjects and then perform cardiovascular, muscular-skeletal and vestibular activities. Furthermore, developing and executing social reintegration strategies are quite complex and depends upon the individual and group level cognitive progression, mental health and behavioral challenges. An additional goal is enhancing the quality of life of the residents by incorporating technology. However, providing care, conducting rehabilitation activities and assisting in ADL is time, labor (therapists or caregivers), and resource expensive work. Therefore, the neurocenters face challenges to maintain the high quality of services, and the staff members experience stress to achieve the rehabilitation targets efficiently. These challenges provided the context to investigate the use of robotics devices equipped with AI systems as a support tool in rehabilitation settings.

3 Robotics in Healthcare - Socially Assistive Robots (SARs)

Researchers have conducted extensive investigation towards developing robotics for healthcare and discussed methods to use commercially available robots in the healthcare field. Kyrarini [26] have broadly categorized robots in the following five ways: care robots, hospital robots, assistive robots, rehabilitation robots and walking assistant robots. According to Wynsbeghe [39], a robot can be named only by the way it is used in the field, regardless of its hardware specification and capabilities. For instance, a robot providing support and assistance in patient care can be termed as a Care Robot [26]. Similarly, Socially Assistive Robots (SAR) collaborate with doctors, physicians or physiotherapists and *"provide assistance and improvement in a wide range of medical applications such as robot-assisted therapies [13, 28], complex-surgical operations [10, 11], or for social engagement with people with special needs like children with autism spectrum disorder (ASD) [1, 8, 9, 33]"* [23]. SARs contribute to assisting the patients physically, cognitively and socially. However, they lack the ability to interact close to human-like, recognize human emotions [30], and perform high degree autonomous tasks [7]. Thus, more intelligent systems and software are required to be developed to communicate and interact naturally.

Due to the fast growth of COVID-19 pandemic, many healthcare organizations aim to adopt health care practices with minimal human contact and physical distances. Many studies point to the employment of SAR in this COVID-19 outbreak for two primary tasks: i) patient monitoring and ii) utilising teleoperation to link clinicians with the patients (who are at high risk of infection transmission) [3, 7, 16]. *"Many researchers focus on the various techniques for the rehabilitation of physical and cognitive impaired people, e.g. [34] establish a virtual reality exposure therapy (VRET) for managing stress reactions. Similarly, [2, 24] develop a BCI system for the extraction of psychological signals of mentally impaired people using electroencephalography (EEG)"* [20] and improved the

Fig. 3.1: Pepper and Nao from Softbank Robotics

quality of life by transmitting these signals to aiding and virtual and augmented reality (VAR) devices. However, one of the limitations of the BCI systems is mounting of the sensors on the head to receive brain signals and transmit to the linked-aiding devices, which restricts natural movement of the subjects under observation [20].

Pepper [31] and Nao [37] from Softbank Robotics as seen in Figure 3.1 are high performance humanoid socially assistive robots (SAR) for research and education purposes with the ability to equip with machine learning or deep trained models to identify human emotions, generate body gestures and contribute in wide range of rehabilitation applications. The Pepper robot is a humanoid robot with four-wheeled base (instead of legs) on which sonar, laser and bumper sensors are mounted. *"There is a 10.1-inch touch display on its torso, and it has a total of twenty Degrees of Freedom (DOF), including six DOF for each hand, two each for the head and hips, one in the knees, and three in the base. The head hosts two RGB cameras, a depth camera, a microphone, and a tactile sensor to perceive the world, and two speakers where the ears would be on a human"* [26]. In contrast, NAO 6 th generation robot is much smaller 22.8 inches humanoid robot with twenty-five DOF, eleven in the legs and fourteen in the upper body, without any display on its torso. Nao posses two RGB cameras, nine touch sensors, four microphones, two infra red emitters and receivers and eight pressure sensors. Both the robots operate on NAOqi operating system, with software Development Kit (SDK) and graphical programming suite (Choreographe) that provides adaptability to be used in diverse applications such as for interactive, educational, navigation, localization and rehabilitative purposes, providing an easy platform for Human Robot Interaction (HRI) [6, 14, 22, 29]. Many researchers have deployed the pepper robot successfully to teach children [37], as a companion for aging people [40]

and as a coach to guide older people with mental disorders through rehabilitation activities [35]. In our studies we have used the Pepper robot for monitoring device, motivational tool and to provide information to staff members in the neuroceters [20].

4 Pepper Robot Deployment in the Neurocenter - A Pilot Study

"A typical robot for health monitoring and improvement needs to receive audio, video or proximity information from its sensors. This information is then processed based on the algorithm that interpret the information into meaningful signals. This is followed with robot action or response for the desired task [33]" [23]. The goal of placing a robot in a neurocenter is to aid the people suffering from neurological disorders presents additional challenges such as non-cooperative behavior, unstable mental conditions, lack of emotional expressions, impaired cognitive and physical activities [18, 20, 21]. SARs are heavily dependent on the audio, video and proximity sensor information for natural interaction. However, residents often have speech inhibitions and physical morbidity thus limiting the performance of these SARs. To deal with these challenges, we have developed a system using deep learning architecture trained on TBI datasets (in three specified scenarios such as cognitive, physical and social communication) for the vision system of the pepper robot. The primary roles for the Pepper intervention are monitoring psychological, physiological signals and promoting social interaction during the therapy sessions. In addition, providing feedback and information to the trainers or staff members to adopt their strategies according to the mood and performance of the subjects.

4.1 Pepper Robot Architecture

Figure 3.2 illustrates the Pepper platform for the field study. The system is built of three main units:
i) Sensory unit , that contains (RGB and depth) visual sensors, (microphones) audio sensors and (laser, sonar and bumper) proximity sensors. This unit acquires and processes visual psychological data and sends it to the deep trained model for emotion and mood classification.
ii) Social signal synthesis unit, that generates the robot gestures and actions based upon the classified emotions and cognitive states. This module enables the robot to perform pre-defined robotic actions based upon the classified emotional signals.
iii) Robot graphical user display (speech, visual and gesture output/actions) unit. The display on the Pepper robot's torso and microphones on head and the Pepper's robot upper body, provide visual, audio and gestural information and feedback to the staff members and subjects under therapy.

4.2 Pepper robot Feedback

The feedback provided by the Pepper robot includes verbal and non-verbal gestures. The feedback categories presented are: i) Monitoring feedback that provides the pool

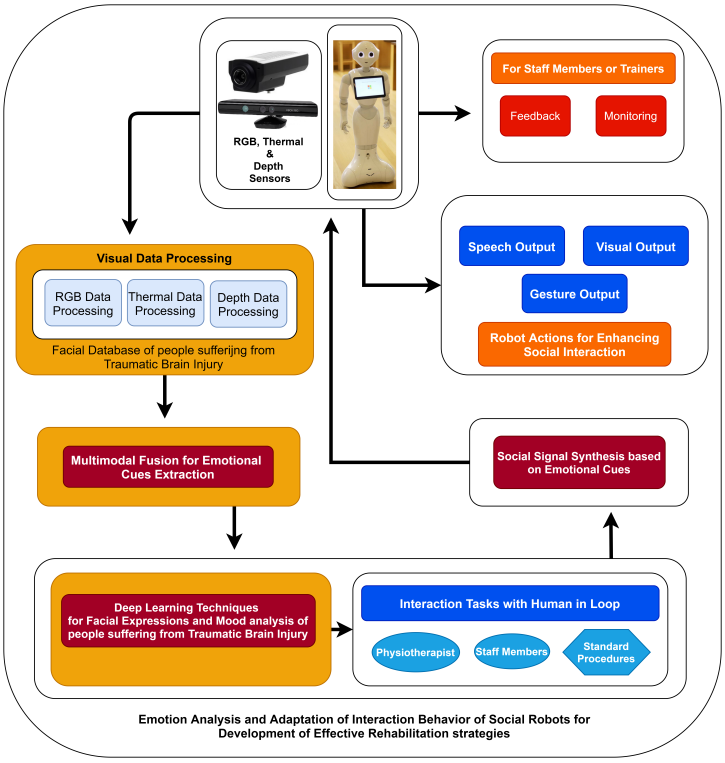


Fig. 3.2: Pepper robot emotion analysis application platform with adaptation of interaction behavior of the Pepper robot for effective rehabilitation strategies

of expression over the entire session and the performance information to the therapist. ii) Emotional and mood study feedback, where the display is used to illustrate the subjects emotional expression and mood through emojis. This feedback is also conveyed through verbal phrases to indicate the mood of the subjects under the therapy sessions. iii) Motivational feedback encourages the patients to conduct more physical activity repetitions.

i) Pepper robot as Monitoring Device

During the field study, Pepper served as a monitoring device for the resident's emotional states in relation to the performance during the rehabilitation activities. This intervention showed a decline in performance when more negative emotional feedback is provided. The Pepper robot monitoring feedback assists staff members and therapists in identifying the emotional states before, during and after the therapy sessions and provides flexibility for adopting the rehabilitation strategies accordingly. Our studies supported the findings of [17], that *"to achieve the best results, it is essential to determine the emotional states of the patients prior to conducting an rehabilitation exercise"* [20]. This could greatly impact in achieving desired rehabilitation goals efficiently and could save therapist time and energy.

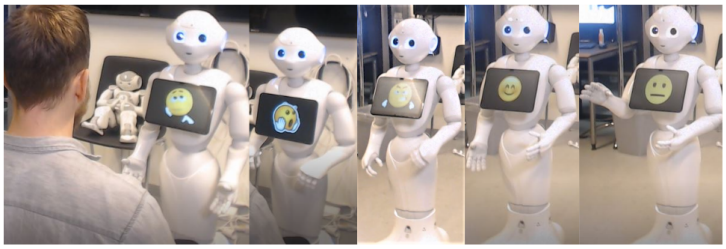


Fig. 3.3: Pepper robot displays emotional and cognitive states feedback

ii) Pepper robot as an Information Tool for Therapists
The pepper robot displays emotional and cognitive states feedback through the verbal and non-verbal gestures as illustrated in Figure 3.3. The verbal gestures are a fixed set of phrases for example *"Great, you look happy, I am happy too, yahoo yahoo yahoo"; "Why do you look so surprised, did i miss something"; "You look angry, please calm down" ; "Don't be sad, I can cheer you up"; "You are quite neutral"* . The non-verbal feedback involves the display of emotions-related emojis on the display of the Pepper robot. The robot is placed in front of the subjects and therapists so that they can visualise this information and adjust accordingly if necessary.

iii) Pepper robot as Motivational Tool
Motivational feedback encourages the patients to conduct more physical activity repetitions. The verbal motivational feedback consists of a set of phrases for example *"Wow, You are doing great!"; "I am so glad you have completed your session"; Give your best buddy"; "Great! you are improving fast"*. It is observed that with the motivational feedback, subjects performed better during the physiotherapy sessions.

In our field study as demonstrated in Figure 3.4 we observed that Pepper audio, visual and gesture feedback impacted positively on the physical rehabilitation but negatively on the cognitive activity. This is due to the fact that Pepper failed to differentiate their focused-emotional reactions from confused or negative emotional expressions [20]. On the other hand, the robot's feedback during cognitive tasks was called distracting by the participants under observation. To counter this problem we have offered Wizard-of-Oz (WoZ) functionality *" to equip the Pepper robot with intellectual cognitive abilities in decision making as well as in creating good relationships with its human user"* [20]. The WoZ feature could aid the therapists and trainers to meet the desired targets during cognitive rehabilitation task and in developing a reliable relationship between robot and human user. However, due to COVID-19 pandemic, we could not conduct the second field study and WoZ feature could not be tested in the neurocenter environment with patients suffering from TBI.

5 Discussion and Conclusion

The study presents the assessment of the Pepper robot intervention for neurorehabilitation. Five patients were evaluated for fifteen sessions performing physical, cognitive

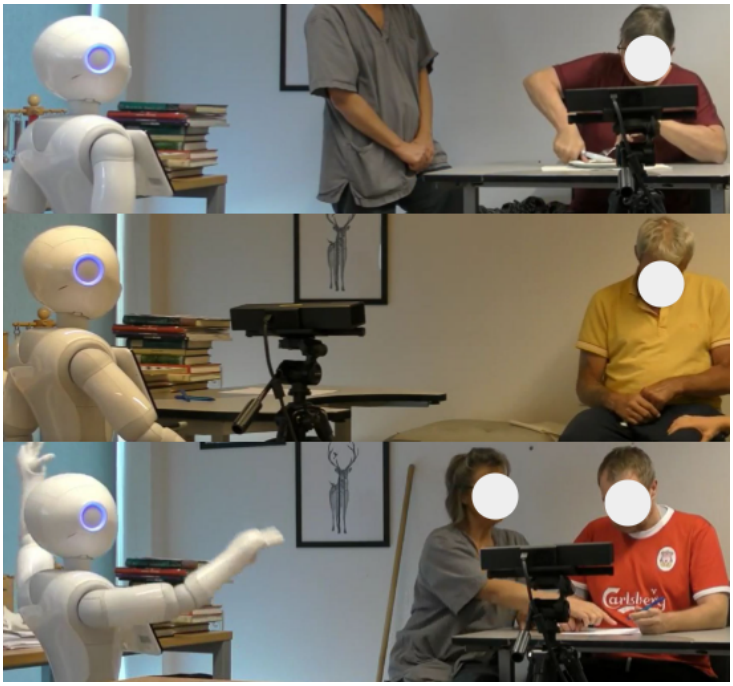


Fig. 3.4: Pepper robot performing field study during physical and cognitive rehabilitation sessions in the neurocenter

and social rehabilitation strategies. The patients performance is compared with conventional and robot-assisted therapy sessions. Overall, the findings of the social robot assisted therapy indicated the positive impact on patients performance and social engagement. The technology also aided the staff members to view the objective recovery and to view the real time monitoring of the emotional states off the patients.

Regarding the acceptability, most of the patients demonstrated openness towards robotic solutions and recognized the beneficial partnership in aiding rehabilitation methods for patients and clinicians. However, the reliability of the system performance is crucial in seamlessly incorporation of robotic platforms in the healthcare environment.

References

[1] J. Abbasi, "In-home robots improve social skills in children with autism," *Jama*, vol. 320, no. 14, pp. 1425–1425, 2018.

[2] V. H. C. d. Albuquerque, R. Damaševičius, N. M. Garcia, P. R. Pinheiro *et al.*, "Brain computer interface systems for neurorobotics: methods and applications," 2017.

[3] L. Aymerich-Franch, "Why it is time to stop ostracizing social robots," *Nature Machine Intelligence*, vol. 2, no. 7, pp. 364–364, 2020.

[4] J. Burgner-Kahrs, D. C. Rucker, and H. Choset, "Continuum robots for medical applications: A survey," *IEEE Transactions on Robotics*, vol. 31, no. 6, pp. 1261–1280, 2015.

[5] C. Caudwell and C. Lacey, "What do home robots want? the ambivalent power of cuteness in robotic relationships," *Convergence*, vol. 26, no. 4, pp. 956–968, 2020.

[6] A. Causo, G. T. Vo, I.-M. Chen, and S. H. Yeo, "Design of robots used as education companion and tutor," in *Robotics and Mechatronics*, S. Zeghloul, M. A. Laribi, and J.-P. Gazeau, Eds. Cham: Springer International Publishing, 2016, pp. 75–84.

[7] N. Céspedes, D. Raigoso, M. Múnera, and C. A. Cifuentes, "Long-term social human-robot interaction for neurorehabilitation: Robots as a tool to support gait therapy in the pandemic," *Frontiers in neurorobotics*, vol. 15, 2021.

[8] P. Chevalier, "Social personalized human-machine interaction for people with autism," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 229–230.

[9] P. Chevalier, J.-C. Martin, B. Isableu, C. Bazile, and A. Tapus, "Impact of sensory preferences of individuals with autism on the recognition of emotions expressed by two robots, an avatar, and a human," *Autonomous Robots*, vol. 41, no. 3, pp. 613–635, 2017.

[10] B. Davies, "Robotic surgery–a personal view of the past, present and future," *International Journal of Advanced Robotic Systems*, vol. 12, no. 5, p. 54, 2015.

[11] S. P. DiMaio and S. E. Salcudean, "Needle steering and motion planning in soft tissues," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 6, pp. 965–974, 2005.

[12] L. D. Evjemo, T. Gjerstad, E. I. Grøtli, and G. Sziebig, "Trends in smart manufacturing: Role of humans and industrial robots in smart factories," *Current Robotics Reports*, vol. 1, no. 2, pp. 35–41, 2020.

[13] D. Feil-Seifer and M. J. Matarić, "Socially assistive robotics," *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 24–31, 2011.

[14] B. Goertzel, J. Mossbridge, E. Monroe, D. Hanson, and G. Yu, "Humanoid robots as agents of human consciousness expansion," *arXiv preprint arXiv:1709.07791*, 2017.

[15] J. Gross, "Interviewing roomba: A posthuman study of humans and robot vacuum cleaners," *Explorations in Media Ecology*, vol. 19, no. 3, pp. 285–297, 2020.

[16] J. E. Hollander and B. G. Carr, "Virtually perfect? telemedicine for covid-19," *New England Journal of Medicine*, vol. 382, no. 18, pp. 1679–1681, 2020.

[17] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 369–389.

[18] ——, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[19] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[20] C. Ilyas, M. Rehm, K. Nasrollahi, Y. Madadi, T. Moeslund, and V. Seydi, "Deep transfer learning in human-robot interaction for cognitive and physical rehabilitation purposes," *Pattern Analysis and Applications*, 2021.

[21] C. M. A. Ilyas, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," *25th IEEE International Conference on Image Processing (ICIP)*, 2018.

[22] C. M. A. Ilyas, K. Rodil, and M. Rehm, "Developing a user-centred communication pad for cognitive and physical impaired people," in *Interactivity, Game Creation, Design, Learning, and Innovation*. Springer, 2019, pp. 124–137.

[23] C. M. A. Ilyas, V. Schmuck, M. A. Haque, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Teaching pepper robot to recognize emotions of traumatic brain injured patients using deep neural networks," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[24] N. M. Krishna, K. Sekaran, A. V. N. Vamsi, G. P. Ghantasala, P. Chandana, S. Kadry, T. Blažauskas, and R. Damaševičius, "An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using eeg signals," *IEEE Access*, vol. 7, pp. 77 905–77 914, 2019.

[25] Y. S. Kwoh, J. Hou, E. A. Jonckheere, and S. Hayati, "A robot with improved absolute positioning accuracy for ct guided stereotactic brain surgery," *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 2, pp. 153–160, 1988.

[26] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon, "A survey of robots in healthcare," *Technologies*, vol. 9, no. 1, p. 8, 2021.

[27] J. Leoste and M. Heidmets, "The impact of educational robots as learning tools on mathematics learning outcomes in basic education," in *Digital Turn in Schools—Research, Policy, Practice*. Springer, 2019, pp. 203–217.

[28] M. J. Matarić, "Socially assistive robotics: Human augmentation versus automation," *Science Robotics*, vol. 2, no. 4, p. eaam5410, 2017.

[29] S. Matsuzoe and F. Tanaka, "How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012, pp. 339–344.

[30] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition

toolkit," in *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2016, pp. 3723–3726.

[31] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[32] K. Rodil, M. Rehm, and A. L. Krummheuer, "Co-designing social robots with cognitively impaired citizens," in *The 10th Nordic Conference on Human-Computer InteractionNordic Conference on Human-Computer Interaction*. Association for Computing Machinery, 2018.

[33] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, p. eaao6760, 2018.

[34] J. Šalkevicius, R. Damaševičius, R. Maskeliunas, and I. Laukienė, "Anxiety level recognition for virtual reality therapy system using physiological signals," *Electronics*, vol. 8, no. 9, p. 1039, 2019.

[35] M. Sato, Y. Yasuhara, K. Osaka, H. Ito, M. J. S. Dino, I. L. Ong, Y. Zhao, and T. Tanioka, "Rehabilitation care with pepper humanoid robot: A qualitative case study of older patients with schizophrenia and/or dementia in japan," *Enfermeria clinica*, vol. 30, pp. 32–36, 2020.

[36] J. Smids, S. Nyholm, and H. Berkers, "Robots in the workplace: a threat to—or opportunity for—meaningful work?" *Philosophy & Technology*, vol. 33, no. 3, pp. 503–522, 2020.

[37] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi, "Pepper learns together with children: Development of an educational application," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 270–275.

[38] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer handbook of robotics*. Springer, 2016, pp. 1657–1684.

[39] A. Van Wynsberghe, "Designing robots for care: Care centered value-sensitive design," *Science and engineering ethics*, vol. 19, no. 2, pp. 407–433, 2013.

[40] C.-Y. Yang, M.-J. Lu, S.-H. Tseng, and L.-C. Fu, "A companion robot for daily care of elders based on homeostasis," in *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. IEEE, 2017, pp. 1401–1406.

Chapter 4

Conclusion and Future Research

This thesis contributes to the development of the methods and models to improve the emotion recognition system for people with neurological impairments, specifically with traumatic brain injury (TBI). In this dissertation, four main contributions have been made towards building robust and high-performance emotion recognition system using: i) newly developed multimodal database of people suffering with TBI, ii) developing and optimizing deep learning architecture based on CNN and LSTM and deep transfer learning for facial features training to enhance the performance of facial emotion classification, iii) fusing RGB and thermal modalities to maximize the emotion and mood classification, iv) robotic application of emotion recognition system to facilitate staff members and people suffering with TBI for cognitive and physical rehabilitation purposes in the neurocenters.

Four main studies are performed to address the aforementioned challenges and improve the emotion classification for the people suffering from the brain injury. The first study involves a comprehensive database establishment from the residents of neurocenters suffering from TBI in three different scenarios that are physical, cognitive, and social rehabilitation scenarios. This study evaluates the facial features variations due to paralysis, physical and cognitive inhibition in the above-mentioned scenarios in a natural and unconstrained environment. This study also includes various pre-processing techniques like face-frontalisation, data augmentation and face quality assessment methods to ensure good quality of data to be processed in the next phases.

The second study presents a deep learning architecture with a linear combination of CNN and LSTM to use spatial and temporal information to identify emotions and cognitive states. This system is trained on TBI dataset and also tested on public datasets like CK+. This study signifies the importance of TBI-database for the improvement of emotion recognition system for people suffering from TBI. This study also explores the use of transfer learning techniques to take advantage of feature learning from larger identity data and fine-tuning to TBI-database.

The third study elucidates the fusion of RGB and thermal facial images. The proposed framework is based on CNN-LSTM network for the fusion of spatio-temporal components of RGB and thermal signals in two different approaches: early fusion and late fusion. The study evaluates and compares the performance of the two multimodal fusion models on TBI dataset. In this study, facial and body gestures are also fused. The study explores and evaluates the performance of early, late, and compact bilinear pooling fusion [6] to enhance the emotion recognition accuracy on FABO datasets.
The fourth study presents the robotic application of emotion recognition to monitor the emotions of people suffering from TBI and to facilitates the staff members for effective rehabilitation purposes. This study explores the intervention of the Pepper robot in the neurocenter to interact with staff members and residents with TBI and evaluates the robotic gesture synthesis on the basis of emotional cues.

This chapter offers an overarching conclusion by presenting the summary of achievements from the above-mentioned four studies and introducing the scope of robotic application of emotion recognition for future studies. Furthermore, a discussion on the limitation and strengths of the research is outlined, as well as further research and speculations regarding the status of robotic therapeutic rehabilitation is offered.

1 Summary of Achievements

1.1 Facial Dataset of people suffering from TBI

Since there is no public dataset of facial features of people suffering from neurological disorders, classifying emotions through faces of healthy people that are mostly available as public facial datasets can lead to inaccurate emotion recognition. People who acquired a brain damage exhibit abnormal facial expressions due to impaired movements of facial muscles. Therefore, for accurate facial emotion recognition of people with traumatic brain injury, it is essential to collect the database of such subjects who have suffered brain injury. In this dissertation, a new facial database has been offered, called TBI-database, including data from 11 subjects, captured with three different sensors, RGB, thermal and Depth. This database is collected in more than 30 sessions in collaborating neurocenters with each subject performing cognitive, physio and social communication activities. This study comprehensively viewed the various data collection strategies in the natural environment and discussed how to overcome the limitations associated with people with brain injury. We proposed various framework along with physiotherapists, psychologists and staff members within the set of possible solutions to improve the facial data provided with a given measure of quality to propose an optimal solution. The results confirmed that the proposed framework effectively improves the quality of data to identify the emotion through facial features. Details of this database has been presented in the Chapter 2, and relevant articles [1–4] This database could not be published due to privacy issues of the subjects, but this database has contributed in all publications made by the author.

1.2 Deep Learning architecture of CNN and LSTM to improve the Emotion Classification

It is challenging to select an appropriate classifier to effectively recognize emotions in a natural and unconstrained environment. This is due to the fact that facial expression is characterized by subtle variations and movements of facial muscles. In fact, emotional cues comprise of time-series variations in facial features within short periods. To exploit the latent temporal information within facial data and to recognize the emotional signals that are produced in a short period of time, it is essential to employ a classifier that deals with temporal information [7]. Therefore, we have used a framework based on the linear combination of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) to use the spatial and temporal information in the facial images of people suffering from TBI. In addition, we have used the VGG-Face descriptor that is trained over 2.6 million face images, built on VGG-16 network, and discarded the last fully connected layer in the architecture. The results showed that incorporation of spatial and temporal information improved the emotion classification even in challenging scenarios.
Although the performance of CNN + LSTM is promising, we have used the transfer learning techniques to benefit from the vastly available facial datasets to overcome the identity limitations of TBI-dataset. We have used eight databases, four recorded in the lab condition and, other four captured in an uncontrolled environment. The results of this study exhibited that the average accuracy of the algorithm is significantly improved with single-source to single-target transfer learning, meeting the goal of the study. Although this framework is computationally expensive, transfer learning is applied offline, and only the fine-tuned model will be used for real-time applications.

1.3 Multimodal Fusion

This study focused on the fusion of the vision and thermal facial images of people suffering from TBI. One of the methods to fuse modalities is to concatenate features from each modality to build one feature vector. However, this approach fails to incorporate non-linear correlation across each modality, which is essential for accurate emotional recognition across various modalities. In this study, to benefit from the non-linear correlation within and across various modalities, a linear combination of CNN and LSTM is applied in two ways; early fusion and late fusion. On the contrary, to short-term six basic expressions, cognitive states and moods are interpreted, thus offering more socially specific cues. Finding from this study showed that fusion of modalities contributed in performance efficacy, with better emotion classification results with late fusion technique. In addition, we have fused facial features and, upper body features, computed in the frame based and sequence -based manner with bilinear pooling fusion on the FABO dataset that involves element-wise multiplication of input features. We have scaled the combined features into a matrix to reduce the dimensions. Although this method is computationally expensive, but improve the accuracy of emotion classification with temporal information.

1.4 Human Robot Interaction for Social Rehabilitation of Patients suffering from Traumatic Brain Injury

In this context, the study contributed in two ways; first through the implementation of deep-trained model in the Pepper robot application for the emotion recognition of people suffering from TBI; and secondly through the intervention of the Pepper robot and gesture synthesis with the aim to enhance social interaction and rehabilitation process. In the first part, the findings of the study showed that deep-trained model on TBI datasets is essential to deal with the emotional and cognitive states recognition of people with brain injury. We compared our results to healthy datasets like CK+ and also evaluated the Pepper-robot emotion recognition model to the TBI datasets. The study showed that to read the subtle emotional changes of the people with injury, it is essential to have a dedicated emotional classifier that could differentiate with challenging emotional states as compared to healthy people. Similarly, our study also supports the implementation of TBI-trained emotional classifier in the Pepper robot rather than using its built-in model. The details of this study have been presented in the [4, 5].

With regards to the Pepper robot gesture synthesis in response to emotion recognition, we conducted the field study and analyzed the results in the following ways:

Pepper Robot as Monitoring Agent
We have conducted the field study in the Danish Neurocenter, where people who have acquired the brain injury are cared for and trained with the activities of daily living (ADL) with the help of staff members, trainers, and physiotherapists. We have introduced the Pepper robot equipped with a customized emotional recognition model to assist the rehabilitation process in three specified scenarios that are cognitive, physical and social rehabilitation. The study provided evidence that intervention of Pepper robot in the Neurocenter to interact with the people with brain injury could be used as a monitoring agent. In our field study data, it is observed that subjects under the rehabilitation process exhibited more mistakes when they are not happy or in a negative mood. Therefore, subjects failed to reach the rehabilitation targets in such emotional conditions. Findings from the study supported that "*it is essential to determine the emotional states of the patients prior to conducting any rehabilitation exercise* " [4]. " *For this purpose, the Pepper robot intervention facilitated the staff members and therapists to determine the emotional states before, during and after the rehabilitation tasks*" and adapt their strategies to meet the goals effectively [4].

To deal with the negative emotions, Pepper robot interacted with subjects by audio, video and gestures synthesis according to the emotional cues. Results from this study showed positive impact on the physical rehabilitation of the subjects. However, this Pepper intervention impacted negatively on cognitive rehabilitation, that was countered by introduction of Wizard-of-Oz (WoZ) functionality.

Pepper Robot as Feedback Agent
In the field study, the Pepper robot recorded the data of the subjects and evaluated the emotions over entire therapy sessions. This provided with the pool of expressions

evaluation against the execution of physical and cognitive tasks. It is observed that during the physiotherapy sessions the Pepper acted as "motivator" for the subjects that resulted in encouragement to perform more repetition of physical tasks. However, for the staff members or trainers, the Pepper acted as "Feedback Agent" that is exhibited through the Pepper audio, video and gesture output and recorded data evaluations to the tasks performance. This study analyzed the emotion expressions and tasks executions with information about the subject-activity-engagement and attention time-span. This feedback aided the therapists and trainers to reflect at their strategies and adopt according to performance of the subjects maintaining the interest of the subjects under observation.

2 Limitations and Future Research

The overarching aim of this thesis was to conduct and improve the emotion-based investigation for the people suffering from neurological disorders using multitude technology and framework. Several algorithms including novel and extensions of existing framework were proposed; including a deep learning architecture based on the combination of CNN and LSTM network to explore the spatio-temporal features of facial datasets to built a robust and reliable emotion classification model. In addition, the Pepper robot integrated with emotion classification model have been introduced to facilitate the process of rehabilitation in the neurocenters. While the findings of this research is encouraging and state-of-the-art results have been published, a number of limitations should be acknowledged.

Although new database with facial features of neurological impaired people has been established as discussed in Chapter 2, it suffers from limited identity data. This database is comprised of eleven subjects, therefore, for future work it is invaluable to increasing number of subjects and train models on larger datasets to improve the classification performance accordingly.

In our second study, we discussed the extraction of the visual features by learning on the large TBI/database through use of CNN and determined the relationship between the transformation of facial expressions in image sequences with the use of LSTM network. In addition, the performance of the emotion recognition system is improved by the use of transfer learning techniques. However, training a network through CNN-LSTM is difficult and computationally expensive. For real-time implementation tensor flow light could be used. Moreover, for effective transfer learning, it is required to have specific labelled dataset for every task. For future research, further investigation on domain adaptation techniques suitable for CNN-LSTM is recommended.

We have combined the visible and thermal features for emotion recognition and it resulted in higher performance. However, it will interesting to investigate the incorporation of depth and audio features towards the improvement in the emotion

recognition.

In our final study, we have offered the Pepper robot intervention to monitor the emotions of the people suffering from the brain injury in the neurocenter for rehabilitation and assistive purposes. In this field study, we have the following findings:

- Pepper robot intervention resulted in the increased number of repetition of physiotherapy activity. However, it will be interesting to determine the same pattern in long-term studies.
- A big challenge in robot assisted rehabilitation is the lack of adherence to the recommended therapy treatments due to decreased compliance and improved treatment outcome. In future research, motivation towards robotic therapy could be enhanced by the installation of more engaging interface with the relevant synthesis of gestures through the body, voice and display of the Pepper robot.
- Although the Pepper robot intervention in the neurocenter produced encouraging results, it is not a part of standard care in the most facilities due to high installation cost. Substantial efforts are being made to develop and implement the low-cost visual-robotic-tools to mitigate the challenging therapy process.
- Due to the COVID-19 pandemic we could not perform the second field study with the Pepper robot. A further investigation is required to analyse the short-term and long-term robotic interaction with the people suffering from the brain injury and outcome of this interaction in relation to desired goals.

References

[1] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 369–389.

[2] ——, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[3] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[4] C. Ilyas, M. Rehm, K. Nasrollahi, Y. Madadi, T. Moeslund, and V. Seydi, "Deep transfer learning in human-robot interaction for cognitive and physical rehabilitation purposes," *Pattern Analysis and Applications*, 2021.

[5] C. M. A. Ilyas, V. Schmuck, M. A. Haque, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Teaching pepper robot to recognize emotions of traumatic brain injured patients using deep neural networks," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[6] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnns for fine-grained visual recognition," in *Transactions of Pattern Analysis and Machine Intelligence (PAMI)*, 2017.

[7] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2015.

Part II

Emotion Recognition

Paper A

Facial Expression Recognition for Traumatic Brain Injured Patients

Chaudhary Muhammad Aqdus Ilyas, Mohammad A. Haque, Matthias Rehm, Kamal Nasrollahi and Thomas B. Moeslund

The paper has been published in the
Proceedings of 13th International Conference on Computer Vision Theory and Applications (VISAPP 2018) Vol. (4), pp. 522–530, 2018.

The layout has been revised.

Abstract

In this paper, we investigate the issues associated with facial expression recognition of Traumatic Brain Insured (TBI) patients in a realistic scenario. These patients have restricted or limited muscle movements with reduced facial expressions along with non-cooperative behavior, impaired reasoning and inappropriate responses. All these factors make automatic understanding of their expressions more complex. While the existing facial expression recognition systems showed high accuracy by taking data from healthy subjects, their performance is yet to be proved for real TBI patient data by considering the aforementioned challenges. To deal with this, we devised scenarios for data collection from the real TBI patients, collected data which is very challenging to process, devised effective way of data preprocessing so that good quality faces can be extracted from the patients facial video for expression analysis, and finally, employed a state-of-the-art deep learning framework to exploit spatio-temporal information of facial video frames in expression analysis. The experimental results confirms the difficulty in processing real TBI patients data, while showing that better face quality ensures better performance in this case.

1 Introduction

Facial expression is one of the main sources of communication for human emotions as approximately 55 percent of human communication is happened through facial expressions [1]. Computer vision techniques have been developed to extract facial features and use them for different purposes [2] [3], for example for assessing, mental states [4] [5], health indicators [6], and various physiological parameters like heartbeat rate, fatigue, blood pressure and respiratory rate [7]. Among these, automatic detection of facial expression is subject of high importance due to its applications in many fields such as in biometrics, forensics, medical diagnosis, monitoring, defence and surveillance [8] [2] [9] [4] [5] [6] [10] [11]. Therefore, researchers are putting great emphasis on development of accurate and robust Facial Expression Recognition (FER) systems. A vast body of literature has been produced on this topic in the past decade.

The existing FER systems can be broadly categorized according to their feature extraction methods [12] and the used classification techniques. Most widely used methods for facial feature extraction are: geometric features based methods, appearance based methods and hybrid ones [13] [14]. Geometry based feature extraction methods use geometric shape and position of the facial parts like lips, nose, eyebrows and mouth, with temporal information such as the movement of facial features points from the previous frame to the current frame [15] [16]. Geometric features are resistant to illumination variation so non-frontal head postures can be handled by processing the figure to frontal head pose to extract the features by measuring distance of fiducial points [17] [18]. Appearance based methods were employed by researchers by using texture information of facial images [19] [20]. In hybrid feature extraction methods both geometric as well as appearance based approaches are deployed for facial image representation [17].

FER systems can be further divided on the basis of classification approaches of extracted facial features. For example, Ghimire et al., 2016 proposed an approach in

which both appearance and geometric features are used for facial expression recognition and Support vector Machine (SVM) for classification [21]. Researchers in [22] [23] have used Local Binary Pattern (LBP), Histogram of Oriented Gradient (HoG) in [21], Linear Discriminant Analysis (LDA) in [24] [22] [23], wavelets based approaches in [25] [26] [17], Non-Negative Matrix Factorization (NMF) and Discriminant NMF in [27] [28]. Lajevardi and Hussain proposed an investigative analysis on feature extraction and selection models for automatic FER system based on AdaBoost algorithm followed by Gabor filters, log Gabor filters, LBP and higher-order local autocorrelation (HLAC), which is then further modified by applying HLAC-like features (HLACLF) [29]. Similarly [15] proposed a temporal based FER by tracking the facial feature points and classifying them using multi-class AdaBost and SVM. In [17], geometric distance specific fiducial points are determined for FER. Researchers in [25] [6] [13] [30] used SVM for accurate classification; whereas authors in [24] used the Hidden Markov Models. SVM shows better results when facial expressions are recognized from single frame, but in case of sequence of images HMM produce better results. It is not the set rule as some authors have used combination of different techniques and produced results comparable to state of the art methods.

In recent years more and more researchers have moved towards deep learning techniques for fast, accurate and robust FER. Authors in [31] [32] [33] applied Deep Convolution Neural Network (DCNN) for classification of features into expressions and achieved appreciable results. Yoshihara et al. proposed a feature point detection method for qualitative analysis of facial paralysis using DCNN [34]. For initial feature point detection, Active Appearance Model (AAM) is used as an input to DCNN for fine tuning. Deep Belief Network (DBN) is another widely used method for robust FER. Kharghanian et al. [35] used DBN for pain assessment from facial expressions, where features were extracted with the help of Convolution Deep Belief Networks (CDBN) to identify the pain. Like [36], it is tested on the publicly available UNBC McMaster Shoulder Pain database with 95 percentage accuracy. However, these existing methods of FER from healthy people, as used in [22] [13] [35], are not suitable when applied to real patients in a real scenario.

Recently [37] proposed a pain assessment system with FER, where CNN is used to learn facial features from VGG-Faces, then linked to Long Short-Term Memory (LSTM) to take advantage of temporal relations between video frames. This method was further improved by [32] by feeding super-resolved facial frames to the CNN+LSTM architecture. These systems of [37] and [32] work well for extraction of facial expression and its interpretation in form of social signals for healthy people. However, the performance of those systems are yet to be tested on datasets collected on the real patients' scenarios like Traumatic Brain Injured (TBI) patients in a care giving center. This mainly because these patients behavior might be very non-cooperative and non-compliance, and they can have agitation, confusion, loud verbalization, physical aggression, dis-inhibition, impaired reasoning, poor concentration, judgment and mental inflexibility [38]. Brain injured patients may also have reduced expressions such as smiling, laughing, crying, anger or sadness or their responses may be inappropriate. On the contrary, some TBI patients also exhibit extreme responses like sudden tears, anger outbursts or laughter. It's all due to loss of ability to control over emotions to some extents. These raise the questions whether the state of the art

FER systems, like [37] and [32], will be reliable when working with these patients data. The main issue is that these system require facial images that are good quality and well-posed towards the camera. However, due to the mentioned issues the TBI patients can not always face the camera and their facial images are not of good quality with certainty due to for example rapid changes in head pose. To deal with these difficulties, we equip the state of the art FER system of [37] with a Face Quality Assessment (FQA) system that discards most of the faces that are not useful for the FER system and feeds the FER system only with faces that are of better quality compared to the other facial images. We have tested the proposed system on real data of TBI patients which has been collected in a Neurocenter in which these patients are taken care of. To the best of our knowledge, no one has done any previous work on TBI patients to understand their facial expressions using computer vision techniques. Therefore this work presents a novel experience in this regard and opens up notion for enhancing social communication between patients and care givers.

The rest of this paper is organized as follows. Section 4 describes the proposed methodology for facial feature extraction and recognition of expressions. Section 5 presents the results obtained from the experiments. Finally, Section 7 concludes the paper.

2 The Proposed Method

This section describes the architecture of the proposed method for FER analysis in a real patient scenario. The block diagram of the proposed method is illustrated in Figure C.1. Following [37], in the first step, the face is detected from a input video. In order to reduce erroneous detection of face we employ a face alignment approach by detecting facial landmarks. The detected landmarks are tracked and faces are cropped according to the landmark positions. In the next step face quality is assessed by following [39] and only good quality faces are stored in face log. Faces are then fed to a CNN. This network was pre-trained with VGG-16 faces as used by [32] and [37]. These steps of the system are further explained in the following subsections.

2.1 Data Acquisition and Preprocessing

The subjects are filmed by a Axis RGB-Q16 camera with resolution of 1280 x 960 to 160 x 90 pixels at 30fps (frames per second). Then, these images are fed to a facial image acquisition system which consists of three steps: face detection, face quality assessment and face logging. The first step is face detection from the video frames for which we used a well-know method, called VJ (Viola and Jones) face detector [40]. Due to its speed and moderately high accuracy by using Haar-like features we selected this method. This method constructs a classifier with the help of learning algorithm based on AdaBoost which effectively classify the images on the basis of few critical features from large set and discard background regions by cascading. However, it is prone to erroneous detection when face is in low quality in terms of occlusion or pose variation. On the other hand, while most FER databases have near-frontal head poses of good quality images with very less occlusions (no spectacles, hand gestures covering the mouth, etc.), in our case subjects are TBI patients and

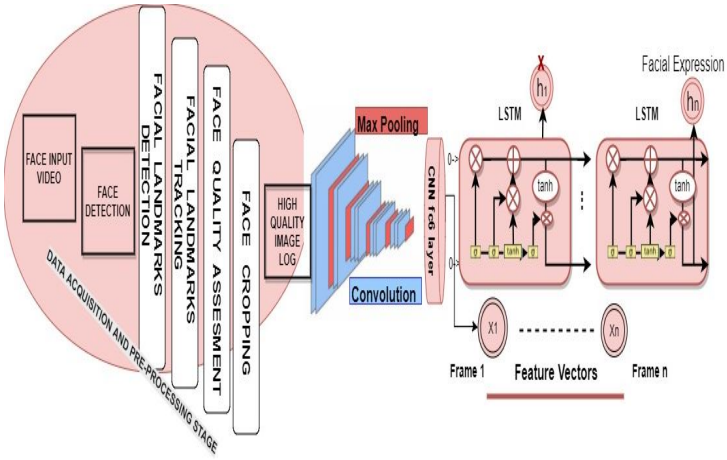


Fig. A.1: Block diagram of Facial Expression Recognition System based on CNN+LSTM model to exploit spatio-temporal information

they are not cooperative enough to ensure good facial data capturing. So there is high possibility of non-frontal view and continuous pose variations, resulting in low quality of images and consequently large amount of miss detected faces as shown in Figure A.2. Moreover, due to inability to recognize and appropriately respond to non-verbal cues TBI patients have feeble response [41]. This in turns increases the complexity of data collection. Thus, instead of detecting face in every single frames of a video, we employ a face alignment method on a properly detected face frame in the video and then track the facial landmarks in the subsequent frames. This reduces the possibility of erroneous detection by VJ in subsequent video frames, as the face is tracked instead of detected again and again in the video sequence.

Face alignment is a process of localization of inner facial structures such as apex of the nose or curve of the eye by using some predefined landmarks that help in better enrolment of the face. Such land-marking also helps in the speedy extraction of geometric structures as well as additional strong local characteristics. Due to advancement in technology, regression based facial land-marking methods have contributed towards the automatic face alignment. One of the most effective approach is the Supervised Decent Method (SDM) [42]. In SDM, 49 facial landmarks are applied around eyes corners, nose line, lips and eye borrows. In addition, SDM uses small optical flow vectors and pixel by pixel neighbourhood measure by avoiding window based point tracing. This provides high computational efficiency, and more stable and precise tracking for long time period of visual frames as demonstrated by [39]. Thus, we employ the SDM based face alignment in the proposed method of FER. The steps of face alignment in a video is shown in Figure A.3. The face is first detected in a video frame by an off-the-shelf face detector (VJ in this case) and then the facial landmarks are identified in that frame. Instead of detecting face in the subsequent video frames, those landmarks are tracked in the subsequent frames. The performance of following the SDM-based approach over mere VJ will be evaluated in the experimental result

section. By using the landmarks, we find the face boundary and then crop the faces. The faces are then forwarded to the next step.

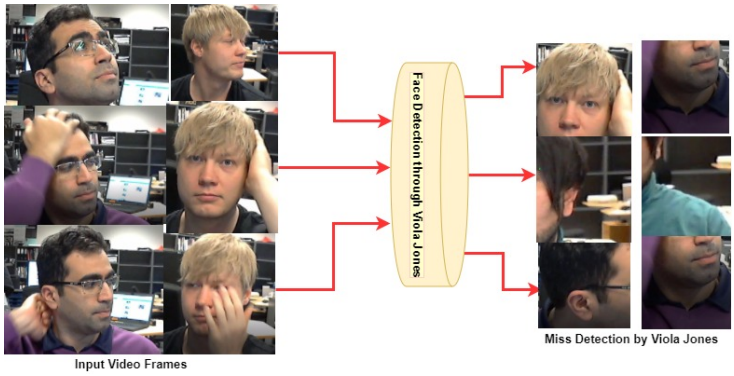


Fig. A.2: Miss detection of faces by VJ face detector due to occlusion or high pose variation

Fig. A.3: Facial landmark identification and tracking in Supervised Descent Method (SDM)

2.2 Face Quality Assessment

System performance for FER is highly dependent on the quality of facial images. In practice for the TBI patient dataset, there is high possibility of non-frontal view of face and continuous pose variations, resulting in low quality of images, even though those faces are tracked by the SDM. Figure A.4 show the case of occluded face (which of course means low quality) for a video sequence where average pixel intensities are varying due to the presence and absence of occlusion over time. To avoid such problems, we employ a FQA technique on the faces cropped after SDM. This is accomplished by measuring some face quality matrices like image resolution, sharpness, and face rotation as shown in [43]. Before logging facial frames into final face log for FER, low quality face frames are identified by setting first frame as a reference frame and comparing similarity in the rest of the frames in a particular event of video as follow in [44]. Similarity of frames is calculated by the following equation:

$$S_{Clr} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(\mathbf{A}_{mn} - \overline{A})(\mathbf{B}_{mn} - \overline{B})}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}(\mathbf{A}_{mn} - \overline{A})^2 \sum_{m}^{M}\sum_{n}^{N}(\mathbf{B}_{mn} - \overline{B})^2}} \tag{1}$$

In the above equation A and B are the reference faces whereas $\overline{\mathbb{A}}$ and $\overline{\mathbb{B}}$ are average pixels levels of the current frame. M and N are number of rows and columns in an image matrix. The degree of dissimilarity calculated from the above equation forms the basis for face quality score. The more the dissimilarity the more the possibility of a low quality face.

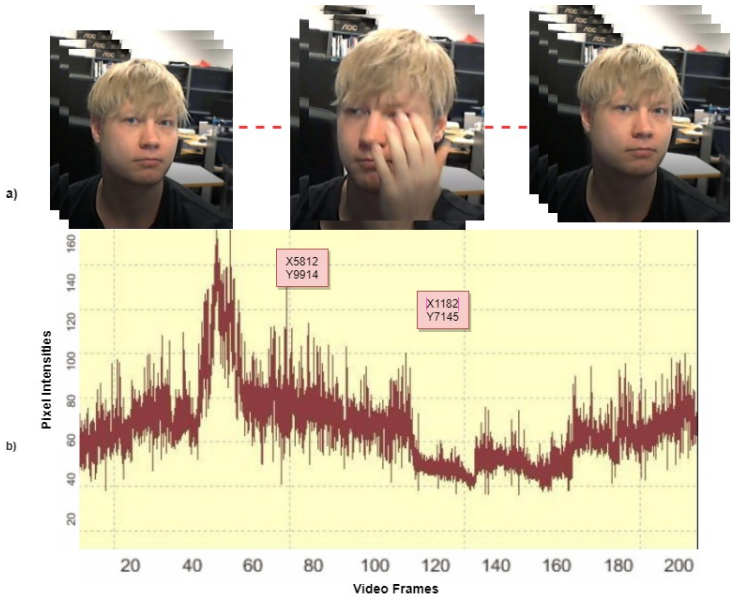


Fig. A.4: Depiction of varying pixels intensities due to the presence and absence of occlusion over time. a) shows the example face frames and b) show the variation in pixel intensities over time

2.3 Face logging

In this step, the faces obtained after SDM tracking are considered along with their associated quality score. If the score is lower than a predefined threshold we simply discard that before logging. Once the quality of face is ensured, images are cropped to a common input size of neural network (224x224 pixels in our experiment) and these are ready to feed to the deep learning architecture.

2.4 The CNN+LSTM based Deep Learning Architecture for FER

Convolutional neural networks are specialized set of neuron networks having multiple layers of input and output that utilizes the local features in image to obtain the visual information. CNN has multiple layers for convolution and padding. A typical 2-Dimensional (2D) CNN takes 2D images as input and considers each image as a $n \times n$ matrix. Generally, parameters of the CNN are randomly initialized and learned by performing gradient descend using a back propagation algorithm. It uses a convolution operator in order to implement a filter vector. The output of the first convolution will be a new image, which will be passed through another convolution by a new filter. This procedure will continue until the most suitable feature vector elements $\{V_1, V_2, ..., V_n\}$ are found. Convolutional layers are normally alternated with another type of layer, called Pooling layer, which function is to reduce the size of the input in order to reduce the spatial dimensions and gaining computational performances and translation invariance [45]. CNN performed remarkably well in facial recognition [46] as well as automatic facial detection [31]. In order to take advantage of its good results for FER we have applied this method on TBI patients data to extract facial features relevant to FER.

In general, CNN deals with images that are isolated. However, in our case we have used the sequences of images in a timely manner and thus, having the notion of using temporal information as well. So to exploit the temporal information associated with facial expression in video, we have used an implementation of Recurrent Neural Network (RNN), that is capable of absorbing the sequential information, called LSTM model from [37]. The LSTM states are controlled by three gates associated with forget (f), input (i), and output (o) states. These gates control the flow of information through the model by using point-wise multiplications and sigmoid functions σ, which bound the information flow between zero and one by the followings:

$$i(t) = \sigma(W_{(x\to i)}x(t) + W_{(h\to i)}h(t-1) + b_{(1\to i)}) \tag{A.1}$$

$$f(t) = \sigma(W_{(x\to f)}x(t) + W(h \to f)h(t-1) + b_{(1\to f)}) \tag{A.2}$$

$$z(t) = tanh(W_{(x\to c)}x(t)) + W_{(h\to c)}h(t-1) + b_{(1\to c)}) \tag{A.3}$$

$$c(t) = f(t)c(t-1) + i(t)z(t), \tag{A.4}$$

$$o(t) = \sigma(W_{(x\to o)}x(t) + W_{(h\to o)}h(t-1) + b_{(1\to o)}) \tag{A.5}$$

$$h(t) = o(t)tanh(c(t)), \tag{A.6}$$

where $z(t)$ is the input to the cell at time t, c is the cell, and h is the output. $W_{(x\to y)}$ are the weights from x to y.

In this paper, we use a combination of CNN and LSTM where CNN extract facial features from the faces logged from the TBI patients video and LSTM find temporal correlation based on those features in temporal setting. A schematic diagram of the

CNN+LSTM is shown in the right hand side of the Figure C.1 and more details can be found in [37]. We used a off-the-shelf fine-tuned version of the VGG-16 CNN model [47] pre-trained with faces for spatial feature extraction. We obtained the features of the `fc7` layer of the CNN (VGG-16) and then use them as input to a the LSTM to exhibit hybrid deep learning performance by CNN+LSTM. The implementation of the CNN+LSTM is available online through [37].

3 Experimental Results

In this section, we first describe the database captured and used during our investigation. We then demonstrate and commented on the results.

3.1 The Database

In order to have experiments for FER on TBI patients data, we require a database. However, to the best of our knowledge, there is no publicly available facial video database from real TBI patients. In establishment of a database, first task was identification of data collection methods. Most of TBI patients have varying ability to identify and respond to non-verbal expression of emotions [41]. After visiting different neurocenters and care-homes where TBI patients are provided rehabilitation facilities around Denmark, and consulting with experts and care-givers who are in direct contact with TBI patients, we have finalized three uniform scenarios for data collection from all the patients under observation. The uniformity in data collection is maintained to have reliable data for future use. Those scenarios are: a) cognitive rehabilitation therapy, b) physiotherapy, and c) social communication with other residents of the neurocenter. In cognitive therapy, a TBI patient plays a game or mind quiz in order to judge how much thinking or cognitive ability a particular subject posses. On the basis of this activity further data elicitation process is organized. In the second activity of physiotherapy, subjects stress level of fatigue is determined. The last activity, where TBI patients have to interact with other patients and care-givers, provides insight about patient ability to give and perceive communication signals.

On contrary to normal people, TBI patients have intolerance, rapid mood swings accompanied by anger or tear bursts, low concentration and impaired facial emotion recognition. Considering these challenges, collection of data, particularly facial videos, is not a trivial task as most of patients do not keep their face positions still. Even if they do so, it is still not easy to understand their emotions for some other problems. Mostly they have sad or depressed emotions after post traumatic life. However, experts who are dealing with TBI patients over certain period of time are able to annotate the patients emotional status as neutral or normal expression. Another problem is: they get agitated very quickly and so it was big task to involve them in the aforementioned three activities. For this purpose, to have clear and precise emotion recognition, we devised a game in such a way that we intentionally let the patients to win to see their happy expressions. Similarly to have their head posed in front of camera, a tablet displaying emotional scenes, is placed just parallel to camera recording their facial expressions. Similar adjustments are made in other activities during

recording. One interesting observation is that all the TBI patients have taken deep interest in mind game, and movie or picture illustration regardless of their disability nature. This allows us to collect more neutral, happy and angry expressions. However, we could not collect much expressions of sadness, surprised and fatigue due to non-cooperation, traumatic disabilities and other social and technical issues.

We collected data in multiple phases in a number of sessions. In total we got 539 video sequences (one sequence means one expression event) with variable lengths (1-5 seconds). However, we observe that the data is highly imbalanced as out of 539 events 463 are of neutral expression. In other words, out of approximately 20,000 frames, almost 14000 represents neutral expressions. Among others, 108 events (app. 3300 frames) of happy, 72 events (app. 2200 frames) of angry and very few are other expressions. On other hand, most of them have too much head motions, so making the data even more challenging for further processing.

3.2 Performance Evaluation

In this section, we first demonstrate the impact of employing a SDM-based face aligner and tracker over VJ face detector. Table A.1 shows the amount of erroneous face detection in the video frames. From the results, we observe that FQA removed 2429 erroneous faces out of 27689 while using VJ. It means that 8.67 percentage of the detection were not correct by VJ. On other hand, when FQA technique is employed on faces detected by SDM, 4.46 percentage of the facial frames were not detected correctly as FQA discarded 1128 frames out of 25289 frames. Comparing both results, SDM-based detection by using alignment and tracking provided better accuracy in finding the right faces.

Table A.1: The performance of SDM-based face alignment and tracking to extract faces from the video frames in comparison to basic VJ face detector.

Number of Frames	**VJ**	**SDM**
Total no. of frames	27689	25289
Training frames	22082	20403
Testing frames	5607	4886
Total mis-detection	2429	1128
Percentage Error	**8.67 %**	**4.46 %**

Table B.3 shows the accuracy of FER in terms of AUC for two scenarios while the number of epochs in the LSTM was varying in yielding the results. The epochs of CNN-LSTM system is gradually increased by step of 5, from 5 to 50 keeping other parameters such as RHO, recurrent depth, and drop-out probability constant. From the results we observe that the accuracy of VJ-based CNN-LSTM system is increased with gradual increase in epochs up to 25 epochs. It reached up to level of 76.94 percent at the 25th epoch. At 30th epoch, its value was dropped down to 67.31 percent, but strangely jumped to 75.26 percent in a higher values of epochs.

Table B.4 show the effect of chaning RHO value for three scenarios. From the results we observe that the SDM-based approach reached maximum AUC value of 75.26 percent. RHO value is gradually changed at step of 2, from 1 to 11, means giving more temporal information for FER, while keeping the epochs constant. AUC values showed the VJ-based approach exhibits slightly higher accuracy by increasing temporal information. In contrast, SDM-based approach got the accuracy above 70 percent in all steps with maximum value of 73.38 percent and minimum value of 70.17 percent. Similar uphill trends is observed up to RHO 5 and then a slight decline is observed.

It is clearly evident from the experiment results for TBI patients data, despite of the challenging datasets accuracy of system is increased to certain extent.

Table A.2: AUC results for FER of TBI patients data with gradual increase in epoch values.

Area Under Curve (AUC)		
Epocs Value	**Viola Jones**	**SDM**
10	66.37	69.49
15	69.55	72.03
20	75.42	63.21
25	76.94	72.96
30	67.31	75.26
35	75.76	72.35
40	63.03	73.38
45	67.08	71.81
50	68.63	74.85

Table A.3: AUC results for FER of TBI patients data with gradual increase in RHO values.

RHO Values	**Area Under Curve (AUC)**		
	Full Frames	**Viola Jones**	**SDM**
1	51.43	61.18	70.17
3	54.26	62.08	72.09
5	53.21	63.03	73.38
7	59.27	63.57	72.27
9	57.12	64.5	72.83
11	59.17	63.29	71.09

4 Conclusion

In this paper, we pointed out the rationale about investigating facial expression analyzing system by using data obtained from real TBI patients. The study reveals the challenges associated with real-world scenarios including patients, instead of healthy volunteers used in the previous works. We captured data from TBI patients in a neurocenter, extracted faces from the video frames by employing different methods to find out the effective one. We then fed the cropped faces into a CNN+LSTM based deep learning framework to exploit both spatio-temporal information to detect the patients mental status in terms of facial expressions. The results were demonstrated with different spatio-temporal parameters of the system. The result showed that the facial information obtained from patient is varying in such a way that it is hard to predict the expression with high accuracy. Moreover, we observed strong effect of employing an effective face detection method with face quality assessment for FER. However, as a note for future work, further processing such as face frontalization, larger dataset for training and subject specific knowledge base incorporation might be useful in improving the performance.

References

[1] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 1.2, no. 4, pp. 53–56, 1968.

[2] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, *Face Detection without Bells and Whistles*. Springer International Publishing, 2014, pp. 720–735.

[3] J. Klonovs, M. A. Haque, V. Krueger, K. Nasrollahi, K. Andersen-Ranberg, T. B. Moeslund, and E. G. Spaich, *Monitoring Technology*. Cham: Springer International Publishing, 2016, pp. 49–84.

[4] M. P. Hyett, G. B. Parker, and A. Dhall, *The Utility of Facial Analysis Algorithms in Detecting Melancholia*. Cham: Springer International Publishing, 2016, pp. 359–375.

[5] Y. Chen, *Face Perception in Schizophrenia Spectrum Disorders: Interface Between Cognitive and Social Cognitive Functioning*. Dordrecht: Springer Netherlands, 2011, pp. 111–120.

[6] F. Li, C. Zhao, Z. Xia, Y. Wang, X. Zhou, and G.-Z. Li, "Computer-assisted lip diagnosis on traditional chinese medicine using multi-class support vector machines," *BMC Complementary and Alternative Medicine*, vol. 12, no. 1, p. 127, Aug 2012.

[7] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial video-based detection of physical fatigue for maximal muscle activity," *IET Computer Vision*, vol. 10, no. 4, pp. 323–329, 2016.

[8] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J Pers Soc Psychol*, vol. 17, no. 2, pp. 124–129, Feb 1971.

[9] S. Du and A. M. Martinez, "Compound facial expressions of emotion: from basic research to clinical applications," *Dialogues Clin Neurosci*, vol. 17, no. 4, pp. 443–455, Dec 2015.

[10] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, *Heartbeat Signal from Facial Video for Biometric Recognition*. Cham: Springer International Publishing, 2015, pp. 165–174.

[11] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed. Springer Publishing Company, Incorporated, 2011.

[12] Y. I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, Feb 2001.

[13] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, April 2006.

[14] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, Feb 2014.

[15] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[16] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Constructing facial expression log from video sequences using face quality assessment," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, Jan 2014, pp. 517–525.

[17] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. N. Yanushkevich, "Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, p. 17, Sep 2012.

[18] R. N. Anwar Saeed, Ayoub Al-Hamadi and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in Human-Computer Interaction*, vol. 2014, pp. 1–13, 2014.

[19] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[20] Y. li Tian, "Evaluation of face resolution for expression analysis," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, June 2004, pp. 82–82.

[21] D. Ghimire, J. Lee, Z.-N. Li, and S. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7921–7946, Mar 2017.

[22] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[23] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," *Sensors*, vol. 11, no. 10, pp. 9573–9588, 2011.

[24] M. Z. Uddin and M. M. Hassan, "A depth video-based facial expression recognition system using radon transform, generalized discriminant analysis, and hidden markov model," *Multimedia Tools and Applications*, vol. 74, no. 11, pp. 3675–3690, Jun 2015.

[25] G. Palestra, A. Pettinicchio, M. Del Coco, P. Carcagnì, M. Leo, and C. Distante, *Improved Performance in Facial Expression Recognition Using 32 Geometric Features*. Cham: Springer International Publishing, 2015, pp. 518–528.

[26] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790 – 799, 2014, best of Automatic Face and Gesture Recognition 2013.

[27] G.-J. de Vries, S. Pauws, and M. Biehl, *Facial Expression Recognition Using Learning Vector Quantization*. Cham: Springer International Publishing, 2015, pp. 760–771.

[28] A. Ravichander, S. Vijay, V. Ramaseshan, and S. Natarajan, *Automated Human Facial Expression Recognition Using Extreme Learning Machines*. Cham: Springer International Publishing, 2016, pp. 209–222.

[29] S. Lajevardi and Z. Hussain, "Novel higher-order local autocorrelation-like feature extraction methodology for facial expression recognition," *IET Image Processing*, vol. 4, pp. 114–119(5), April 2010.

[30] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, Jan 2007.

[31] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ser. ICMR '15. ACM, 2015, pp. 643–650.

[32] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*. Cham: Springer International Publishing, 2017, pp. 151–162.

[33] D. Triantafyllidou and A. Tefas, "Face detection based on deep convolutional neural networks exploiting incremental facial part learning," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 3560–3565.

[34] H. Yoshihara, M. Seo, T. H. Ngo, N. Matsushiro, and Y. W. Chen, "Automatic feature point detection using deep convolutional networks for quantitative evaluation of facial paralysis," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Oct 2016, pp. 811–814.

[35] R. Kharghanian, A. Peiravi, and F. Moradi, "Pain detection from facial images using unsupervised feature learning approach," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 419–422.

[36] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain?" in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, Feb 2017, pp. 1–8.

[37] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2017.

[38] M. D. Lauterbach, P. L. Notarangelo, S. J. Nichols, K. S. Lane, and V. E. Koliatsos, "Diagnostic and treatment challenges in traumatic brain injury patients with severe neuropsychiatric symptoms: insights into psychiatric practice," *Neuropsychiatr Dis Treat*, vol. 11, pp. 1601–1607, 2015.

[39] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Quality-aware estimation of facial landmarks in video sequences," in *2015 IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 678–685.

[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[41] J. Bird and R. Parente, *Recognition of nonverbal communication of emotion after traumatic brain injury*, 2014.

[42] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 532–539.

[43] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Aug 2013, pp. 443–448.

[44] R. Irani, K. Nasrollahi, A. Dhall, T. B. Moeslund, and T. Gedeon, "Thermal superpixels for bimodal stress recognition," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.

[45] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–1, 2017.

[46] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[47] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

Paper B

Effective Facial Expression Recognition Through Multimodal Imaging for Traumatic Brain Injured Patient's Rehabilitation

Chaudhary Muhammad Aqdus Ilyas, Mohammad A. Haque, Matthias Rehm, Kamal Nasrollahi and Thomas B. Moeslund

This paper has been published as a book chapter in the *Computer Vision, Imaging and Computer Graphics Theory and Applications. Part of the Communications in Computer and Information Science book series* Springer, (CCIS, volume 997), pp. 369-389, 2019. doi: 10.1007/978-3-030-26756-8$_{1}$8.

The layout has been revised.

Abstract

This article presents the issues related to applying computer vision techniques to identify facial expressions and recognize the mood of Traumatic Brain Injured (TBI) patients in real life scenarios. Many TBI patients face serious problems in communication and activities of daily living. These are due to restricted movement of muscles or paralysis with lesser facial expression along with non-cooperative behaviour, and inappropriate reasoning and reactions. All these aforementioned attributes contribute towards the complexity of the system for the automatic understanding of their emotional expressions. Existing systems for facial expression recognition are highly accurate when tested on healthy people in controlled conditions. However, their performance is not yet verified on the TBI patients in the real environment. In order to test this, we devised a special arrangement to collect data from these patients. Unlike the controlled environment, it was very challenging because these patients have large pose variations, poor attention and concentration with impulsive behaviours. In order to acquire high-quality facial images from videos for facial expression analysis, effective techniques of data preprocessing are applied. The extracted images are then fed to a deep learning architecture based on Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM) network to exploit the spatiotemporal information with 3D face frontalization. RGB and thermal imaging modalities are used and the experimental results show that better quality of facial images and larger database enhance the system performance in facial expressions and mood recognition of TBI patients under natural challenging conditions. The proposed approach hopefully facilitates the physiotherapists, trainers and caregivers to deploy fast rehabilitation activities by knowing the positive mood of the patients.

1 Introduction

Traumatic Brain Injury (TBI) lead to life-long harm to physical, cognitive, emotional and behavioral abilities depending upon the area of the brain damage. For example if frontal lobe is damaged, person lacks skills of planning, organizing, emotional and behavioral control, aggression, problem solving, attention, social skills, flexible thinking, consciousness and hand-eye coordination [1]. Similarly if temporal lobe is impaired, it can lead to complications in memory, recognition of faces, emotions elicitation, sequencing and speaking abilities. Moreover, occipital lobe is controlling the visual functions and parietal lobe is responsible for perception, spatial consciousness, objects manipulations and spelling [2]. Rehabilitation after brain injury is complex, long and expensive process, that can vary greatly depending upon the intensity of the injury and some times can result in permanent disabilities [3]. In America almost one million people suffer from brain injury and almost same number of people suffer in Europe each year, with approximately 4 million people are living globally having long-term disability after TBI [4] [5].

Researchers are putting high emphasis on fast and efficient rehabilitation to lessen the suffering and low quality life of TBI patients. Physiotherapists, trainers and caregiver face severe complications in performing rehabilitation tasks as these patients have limited ability to perceive social signals associated with sudden changes

in behavior including aggression, negative emotions, reduced motor and reasoning skills [6] [7] [8]. Experts, psychologists, trainers and researchers strongly believe that rehabilitation process can be made fast by accurately assessing the emotions of these patients [7] [9]. Researchers are enforcing computer vision (CV) techniques for automatic assessment of mental states [10] [11], monitoring elderly people [12] and measuring various physiological parameters like heartbeat rate, fatigue, blood pressure and respiratory rate [13], in a contactless manner by analyzing facial features [14] [6]. Therefore, researchers are focusing with greater intensity in the development of accurate, reliable and robust facial expression recognition (FER) system. Automatic detection and identification of facial features by utilizing CV techniques are cost and time effective with 24/7 monitoring facility and lesser human assessment errors. Due to this, it has wide range of applications in various fields like monitoring, medical examination, forensics, biometric, defense and surveillance [6].

Most of the current computer vision techniques for Facial Expression Recognition (FER) are working effectively and robustly only on the healthy people in controlled environment. But when these systems are applied on TBI patients in real environment, we have to face unique challenges incurred from data collection, pre- and post-processing, expression recognition and environmental conditions. However, to the best of our knowledge there is neither research on data collection techniques from TBI patients nor public database from real patients for facial expression analysis. Thus, we created a database of TBI patients as in [6] and identified that emotional states of TBI patients are quite different from the healthy people with large imbalance of six common expressions along with higher negative emotional states.

The methods proposed by [15] and [16] perform exceptionally well for FER, and its modeling and structuring as social signals for healthy people in controlled environment. However, these systems demonstrated challenges and complications when applied in real environment on real Traumatic Brain Injured (TBI) patients residing at specialized centers like neuro-centers or care-homes [6]. These challenges are associated with non-cooperative and non-compliance patient's behavior, along with varied level of aggression both verbal and physical, agitation, anxiety, disorientation, disinhibition, improper reasoning, lack of concentration, judgment and mental inflexibility [17]. In addition to that, brain injured patients also suffer with limited facial expressions such as smile, laugh, cry, anger or sadness or they may exhibit disproportionate responses. On the other hand, some TBI patients also displays intense responses like abrupt tears, laughter or anger outbursts. This is due to inability to control emotions due to injury. We tested state of the art FER systems, like [15] and [16], on real data of TBI patients in challenging scenarios with variable environmental conditions and figured out there is need of reliable system for FER with high quality facial images that are collected when patients are well-posed towards the camera [6]. Nevertheless, due to the above mentioned reasons, TBI patients do not pose toward camera thus illustrating very large pose variation with poor quality of facial frames. In [6], we employed Face Quality Assessment (FQA) method prior to deep architecture, to remove low quality and unwanted facial images. We also performed 3D face frontalization to acquire more frontal faces. We also developed unique data collection techniques in uniform scenarios to get reliable data from TBI patients in Neuro-centers. Exploiting facial expressions from TBI patients using computer vision

techniques is not much explored field, with no database of these patients. To the best of our knowledge, we are the pioneer in developing database of these patients to understand their facial expressions and analyzing mood for rehabilitation purposes.

This article aims to provide the solution for TBI individuals by developing new tool and extended database for determining and monitoring facial expressions.It also presents unique experience in enhancing communication with patients and care givers, as well as extraction of physiological and psychological signals and interpreting them as social signals. In this article, we compare the results for six expressions as well as classify the facial characteristics into positive and negative states [14]. Experts and psychologists have annotated our collected data and featured negative expressions as fear, disgust, anger, sad, stress and fatigue. Some patients exhibits unique expressions like lip trembling, teeth grinding and frequent eye blinking, which have also characterized as negative expression by the experts [18]. On the contrary, positive expressions have featured as laugh, smile, surprise and few other unique neutral expressions. It is seen that in case of TBI patients, negative expressions are much more abundant then positive expressions during the data collection sessions. With the help of experts, trainers, physiologists, psychologists and caregivers, we determined three uniform scenarios for reliable data collection of TBI patients. Details are explained in section 3.1. We have employed a linear cascading of a Convolutional Neural Network (CNN) and a Long Short Term Memory (LSTM) network [16], with Face Quality Assessment (FQA) and 3D face frontalization on the facial images obtained through RGB and thermal sensors with early and feature level late fusion techniques. Unlike [15, 16], Our approach addressed additional challenges of non-frontal faces, less cooperative and aggressive subjects, high occlusion, low quality of images that required a lot of preprocessing before feeding into system and varied expressions from normal/healthy people. We have also extended the database with more subjects and more effective pre-processing techniques like faster facial landmark detector with D-LIB and 3D-face frontalization than our previous methods in [6, 14]. Experimental results acquired by utilizing deep learning architecture, demonstrated that RGB and thermal modalities in different fusion states assist each other on classifying patients mental states accurately.

The rest of the paper is organized as follows. Next section will describe the related work on FER systems. Section 3.1 describes the creation of the new extended database including camera specifications, data collection arrangements and pre-processing techniques. Section 4 provides the methodology proposed for facial feature extraction and expression recognition. Section 5 demonstrates the results achieved from experimentation. Finally, Section 7 concludes the paper.

2 Related Work

Prevailing FER systems can be distinguished generally on the basis of the techniques used for facial features extraction and classification methods [19]. Facial feature extraction methods can be based on: geometric features, appearance based methods and hybrid ones [20, 21].

- Geometric feature extraction methods make use of geometric shape and posi-

tion of facial components like nose, cheeks, eyebrows, lips, mouth, chin, related to time sequenced information of movement of these salient features. Facial characters movement is analyzed from the previous frame to the current frame [22, 23]. Geometric features are immune to lightning condition fluctuation, that gives flexibility to deal with non-frontal head positions by altering the figure to frontal head pose to extract the features by measuring distance of fixed reference points [24, 25]. Researchers applied effective shape models by using 58 facial landmarks like Pantic et. al. [20].

- Facial feature extraction methods based on appearance utilize the characteristics of a surface of face like skin texture, wrinkles, bulges and furrows. It is not resistant to illumination variation [26, 27].
- Facial feature extraction methods that make use of both geometric features as well as characteristics of surfaces fall in the category of hybrid methods [24]. Hybrid techniques have produced the best results in the development of automatic FER systems.

We can further distinguish the facial expression recognition (FER) systems on the basis of the classification approaches. Since last decade more and more efforts are converged towards deep learning approaches due to fast computational powers and state of art performances. As mentioned in [14] deep learning architecture involving Convolutional Neural Networks (CNN) outperformed traditional methods and provided state of art results for face recognition [28–30], facial expressions recognition [15, 16, 31–37] and emotional states identification [38–41]. These newer approaches like CNN learn the features from the image data for aforementioned computer vision problems, unlike traditional machine learning approaches those use handcrafted features. Handcrafted features such as Local Binary Pattern (LBP), Support vectors, SIFT, Histogram of Oriented Gradient (HoG), Linear Discriminant Analysis (LDA), Non-Negative Matrix Factorization (NMF) and Discriminant NMF and Local Quantized Pattern (LPQ) applied in [42–49]. Although their computational costs are low, CNN-based deep neural networks surpassed them in accuracy. This is because handcrafted features are accompanied with unintended features that have no or less impact on classification. Similarly as these features made by human experts, so not all possible cases are included for features classification. Due to modern advancement in computation devices and invention of many-core GPUs, more and more research is focused around CNNs that has illustrated remarkable success for classification challenges [16, 50–53]. The major advantage of deep learning methods over common machine learning models is the simultaneous performance of feature extraction and classification. Moreover, deep learning methods apply iterative approach for feature extraction and optimize error by back propagation, thus resulting in those important features that human experts can miss while handcrafting features. CNNs are very good at feature learning through training datasets.

Authors in [31], applied deep CNN with Support Vector Machines (SVM) and won first prize in 2013-FER competition. Liu [33] accomplished three tasks- feature learning, feature selection and classification in a consolidated manner and outperformed other methods in extracting extremely complex features from facial images through Boosted Deep Belief Networks (BDBN). The problem of linear feature selections in previous method is addressed by [36] through DBN models. In 2015, Yu

and Zang [34] demonstrated their work for Emotion recognition in Wild challenge for FER based on static images. They have employed multiple deep CNN where each network is randomly initialized thus reduced likelihood and hinge loss, resulting in significantly exceeding the challenge standard criteria. In year 2016, Yoshihara et al. proposed a feature point detection method for qualitative analysis of facial paralysis using DCNN [52]. For initial feature point detection, Active Appearance Model (AAM) is used as an input to DCNN for fine tuning. Kharghanian et al. [53] used DBN for pain assessment from facial expressions, where features were extracted with the help of Convolution Deep Belief Networks (CDBN) to identify the pain. They have further explored Deep Belief Network (DBN) for robust FER. Rodriguez et. al. [15], in 2017 exploited the temporal information by linking Long Short Term Memory (LSTM) with CNN fine tuned with features from VGG-Faces. Their method was boosted by [16] through involvement of deep CNN for fast features extraction and categorization into facial appearances and reinforcing the CNN+LSTM system with super-resolved facial images.

3 THE PROPOSED METHOD

This section describes the main steps of the proposed methods for FER analysis of TBI patients in real challenging scenarios. The block diagram of the proposed method is illustrated in Figure B.1. First step is face detection from input video streams like [15], followed by face alignment by landmark identification to reduce erroneous face detection. These detected landmarks are tracked and then faces are cropped according to the landmark positions. In the next step face quality is assessed by following [54] and only good quality faces are stored in face log. Faces are then fed to a CNN. This network was pre-trained with VGG-16 faces as used by [16] and [15]. These steps of the system are further explained in the following subsections.

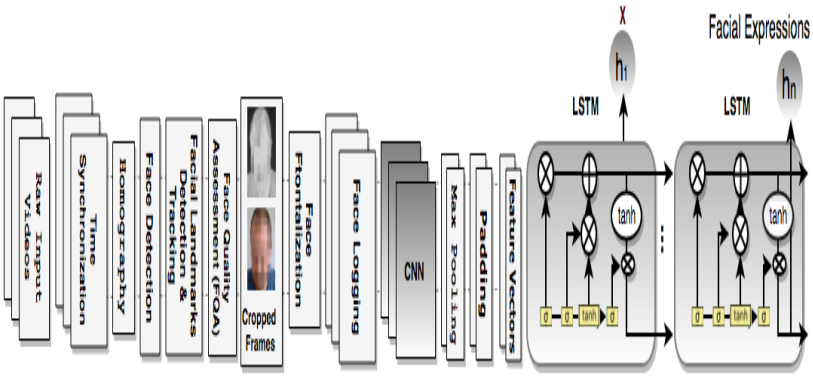


Fig. B.1: Deep learning architecture based upon CNN+LSTM with pre-processing algorithm for Facial Expression Recognition (FER).

3.1 Creating TBI Patient Database

TBI patients have mild to severe injury, accompanied with paralysis, coordination and speech inhibition, and higher level of emotional instability. For facial expression analysis, data is collected from patients who lived in neuro-center for at least 10 weeks prior to commence the data collection so that trainers or care givers can understand their different mental states. This helped in accurate annotation of the data. Physiotherapist, trainers and caregivers devised the rehabilitation strategies by accessing their health indicators, and neuro-psychological and cognitive test results [55]. Data is collected from 9 TBI patients in three pre-defined rehabilitation scenarios in two imaging modalities: RGB and Thermal. These special scenarios are selected with the help of experts, trainers, physiotherapist and care givers by considering the reliability of data as well as disability of patients due nature of injury as described in TableB.1. These scenarios are: 1) Cognitive Rehabilitation, 2) Physical Rehabilitation, and 3) Social Rehabilitation. These are described below.

Cognitive Rehabilitation Scenario

In this scenario, data is collected while patient is performing activities to train the patient's ability to understand particular information and perform function accordingly. Experts perform set of repetitive activities with gradual increase in complexity to asses the memory, attention, visual perception, communication, problem solving and learning skills [56, 57]. In neuro centers, aforementioned task is accomplished by use of calenders, memory devices, drawing clocks, playing quizzes and games, reading or listening books or music, watching movies or other visual aids. Subjects are also given specific tasks like placing room keys at fixed places, telling their daily routine and activities, setting deadlines or time slots for their favourite tasks. These activities are tailored to patients requirements as it is observed while performing aforementioned tasks patients have large pose variations, attention inhibition, less frontal facial pose, emotional instability and aggravated aggression. Different strategies are adopted to enrich the attention and memory of the patient, particularly it is make sure when any subject tells his story or daily routine his or her face must face the camera by placing a mirror just behind the camera and asked them to visualize themselves. Moreover quiz questions, time clocks, calenders, movies and etc. are displayed over tablet placed next to camera, ensuring more frontal images.

Physical Rehabilitation Scenario

In this mode of data collection, patients are performing activities of physical rehabilitation to assess the functionality of sensory motor neurons. Depending upon the nature of the stroke, muscle movements are reduced or ceased. Physiotherapists conducts cardiovascular, skeletal-muscular and vestigial activities to assess the activity tolerance, muscle-action coordination and postural control. These activities are performed through mild walk or running, cycling, push-ups, arms raise, hands or neck moves and other similar activities depending upon a particular subject disability. These physical exercises are modified to have better and reliable facial and upper body data. It is observed that when subjects walk, cycle, or perform similar physical

Table B.1: Database Of TBI patients with Activity Participation

Subjects	Number of Sessions	Activities Participated		
		Cognitive	Social Comm	Physiotherapy
Subject A	7	Y	Y	Y
Subject B	5	Y	Y	Y
Subject C	5	Y	Y	Y
Subject D	7	Y	X	Y
Subject E	3	Y	X	X
Subject F	4	Y	Y	Y
Subject G	3	X	Y	Y
Subject H	6	Y	Y	Y
Subject H	5	Y	X	Y

activity, pose varied largely resulting in very less usable data. To avoid such problems, patients are asked to cycle over stationary bike while keeping their upper body still as much as possible and visualize themselves in specific camera-mirror arrangement as described in previous rehabilitation scenario. Similarly hand pressers, leg raises, walking and other tasks are performed.

Social Rehabilitation Scenario

TBI individuals face severe complexity in social integration due to behavioral and cognitive malfunctions. In this scenario, data is collected while TBI patients are either eating, playing music or discussing or sharing stories, and playing cards or console games in a group of at least 4 or more people. Social communication and integration strategies are also modified according to need of the patients and to have good quality of data. Best results were obtained, when subjects were playing games with the help of consoles and cameras are adjusted next to monitor or screens. It is observed that clear variations in expressions are recorded with changes in the game situations such as happy faces are captured when subjects were winning, sad expression while losing, tense look in difficult situation, even angry looks were observed when cheats are applied in the game.

3.2 Data Acquisition and Preprocessing

Ilyas et.al. [6] established the TBI database with only RGB images where TBI subjects are filmed by Axis RGB-Q16 camera with the resolution of 1280 x 960 to 160 x 90 pixels at 30fps (frames per second) in aforementioned specialized scenarios with tailored techniques. Along with RGB, in this work we have obtained the thermal images with Axis Thermal-Q1922 camera having 10 mm of focal length. After collecting the raw data from both modalities, time synchronization is achieved by the time stamps in the RGB frames captured with variable frame rate. Furthermore, 8-point homography

estimation is employed for approximate image registration by determining homography matrices from RGB to thermal by [58]. These collected images are passed to facial image acquisition system with the following steps.

Face Detection and Tracking

In [6], face detection is performed by Viola Jones (VJ) algorithm. VJ uses Haar like features and employs a classifier based on AdaBoost algorithm to detect the face and discard background. However, due to large pose variation and non cooperative behaviour of TBI patients, VJ detector misses many faces in the video frames when subject has even small non-frontal pose. To address this issue, we have employed deep face detection [59] that gives the more flexibility to detect face even with large pose variation. Deep face detector able to detect face with minimum confidence of 74.24% even when subject has more than 90 degrees of non frontal pose and challenging illumination conditions. After deep-face detection we employed facial landmarks identification and tracking. We have also employed a face alignment method called Supervised Decent Method (SDM) [60] that tracks the facial landmarks in subsequent frames to capture maximum facial images. SDM helped in better enrolment of the face and fast extraction of geometric structures. This also reduces the miss-detection. In SDM, 49 facial landmarks at the apex of nose, curve of eyes, eye borrows, lips and corner of the face are applied. SDM utilizes the optical flow vectors and pixel by pixel neighbourhood measurement which are resulting in high computational efficiency and precise tracking for longer time by avoiding window based point tracing [54]. After aligning the face, the face boundary is determined by landmarks, followed by face cropping. Another advantage of facial landmarks tracking is the reduction of possible erroneous detection by avoiding face detection in each of the video frames.

Face Quality Assessment and Face Logging

Face quality assessment is performed before lodging the facial frames into face log system as it is seen that even capturing the facial images followed by face tracking, there is still presence of unwanted features in the cropped facial images such as hands or hairs over the face or downward faces. As the performance of the system is greatly dependant upon the quality of facial data so these unwanted and erroneous images must be removed. In our case, data is collected from TBI patients who have non cooperative behaviour with continuous head or face movement, and most of the time these movements are combined with hand motion in front of face or camera. Figure B.2 is demonstrating such cases of occluded faces resulting low quality of cropped facial image. In order to avoid such complications, we have devised a filter that discard the faces of low quality on the basis of pixel intensities, image resolution, sharpness and face rotation as shown in [6]. Low quality facial frames are identified by setting first frame as a standard reference frame and discarding rest of the frames who are not 80% or more similar with the first frame in a particular event of video [61]. Similarity of frames is calculated by the following equation:

$$S_{RBG} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(\mathbf{A}_{mn} - \overline{A})(\mathbf{B}_{mn} - \overline{B})}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}(\mathbf{A}_{mn} - \overline{A})^2 \sum_{m}^{M}\sum_{n}^{N}(\mathbf{B}_{mn} - \overline{B})^2}} \times 100\% \tag{B.1}$$

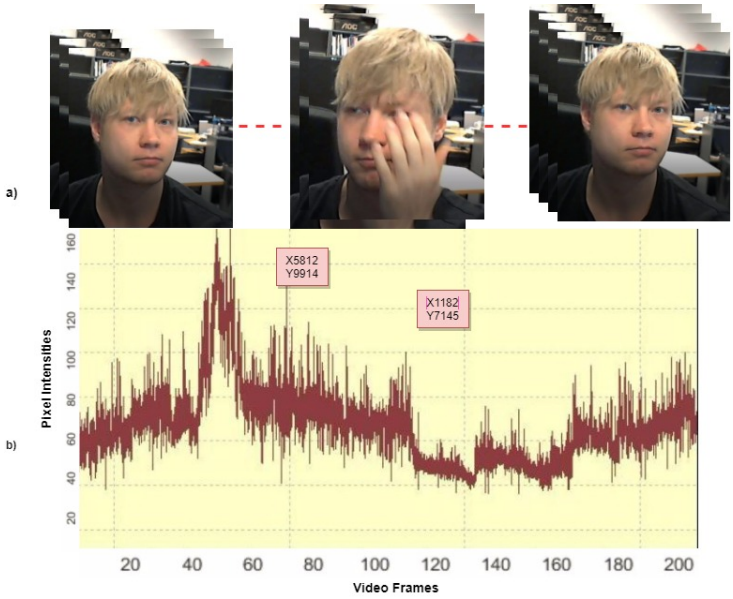


Fig. B.2: Face Quality Assessment overview: (a) Input Image with and without occlusion (b) Varying image pixel intensities due to presence and absence of obstruction in front of the facial image thus aiding in identifying low quality of images to be discarded [6]

In the above equation A and B are the reference faces whereas $\overline{A}$ and $\overline{B}$ are average pixels levels of the current frame. M and N are number of rows and columns in an image matrix. The degree of dissimilarity calculated from the above equation forms the basis for face quality score. The more the dissimilarity the more the possibility of a low quality face. During face logging, when the low quality facial frame in RGB is discarded based on the filter, it's corresponding thermal image is also removed to maintain the synchronization in both modalities. Images are cropped to specific neutral network input size (224x224 pixels in our experiment) after ensuring the best quality of faces.

Face Frontalization

FER is dependant upon the pose of the subject and frontalization can boost the performance of the system many folds. Frontalization is the process of manufacturing frontal facing visual frames showing up in single unconstrained photos [62]. In case of TBI patients with continues and large pose variation, this method has increased the FER accuracy to considerable extent. We have employed the simpler approach of using single, unmodified 3D face, termed as reference face for all the images under observation to produce frontalized sights like [62]. This resulted in better image alignment providing accurate comparison of local facial features of different facial

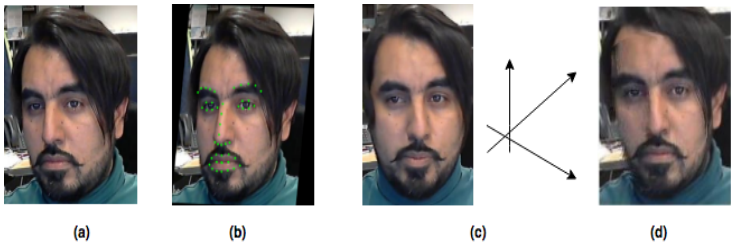


Fig. B.3: Face frontalization process: (a) Query Image, (b) Facial features detection by SDM, (c) Reference face image, (d) Soft symmetry and facial appearance estimation by corresponding symmetric image locations to have frontalized image.

images. Facial features are detected by SDM [60], where pose is estimated by specifying a reference projection matrix C_M consist of intrinsic A_M and extrinsic $[R_M \quad t_M]$ matrices.

$$C_M = A_M * [R_M \quad t_M] \tag{B.2}$$

The extrinsic matrix comprises of rotational matrix R_M and translation vector $t_M]$. Frontal pose is synthesized by taking the transpose of feature points in query image and projecting it on the reference image using geometry of 3D model as seen in figure B.3.As out of plane head rotation leads to less visibility of facial features, this results in occlusion. This is reduced by employing soft symmetry by taking approximation of 3D reference image and single view query image to estimate the visibility in second image. This may result in replication of occlusion as appearances from one side are transferred to another side of the face. In order to avoid this we take the advantage of facial features of aligned images that appear at the same face image locations regardless of the actual shape of the query image.

3.3 Linear Cascading of CNN and LSTM as Deep Learning Architecture for FER

The frontalized facial images obtained by face logging are fed into deep learning architecture composed of CNN and LSTM. CNNs are specialized set of artificial neuron networks with learnable weights and biases. These have multiple input and output layers to analyze the visual information by creating features maps of the image. A schematic diagram demonstrating vital steps in convolutional neural network is represented in the figure B.4. A typical 2-Dimensional (2D) CNN takes 2D images as input and considers each image as a n x n matrix. Generally, parameters of the CNN are randomly initialized and learned by performing gradient descend using a back propagation algorithm. It uses a convolution operator in order to implement a filter vector. The output of the first convolution will be a new image, which will be passed through another convolution by a new filter. This procedure will continue until the most suitable feature vector elements $\{V_1, V_2, ..., V_n\}$ are found. Convolutional layers are normally alternated with another type of layer, called Pooling layer, which function is to reduce the size of the input in order to reduce the spatial dimensions and

gaining computational performances and translation invariant [63]. CNN performed remarkably well in facial recognition [64] as well as automatic facial detection [50]. In order to take advantage of its good results for FER we have applied this method on TBI patients data to extract facial features relevant to FER.

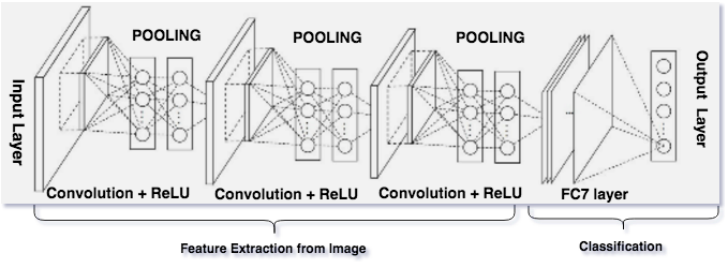


Fig. B.4: Convolution Neural Network (CNN) working paradigm with input, convolution, pooling layers and feature vectors. Adopted from [65]

In general, like any other neural network, CNN deals with images that are isolated. However, in our case we are dealing with the events in a video that happened in time sequential approach so providing the notion of using temporal information. In order to utilize the temporal information associated with facial expression in video, we have used a special kind of Recurrent Neural Network (RNN) that is capable of absorbing the sequential information as well as learning long-term dependencies, called LSTM model from [15]. The LSTM states are controlled by three gates associated with forget (f), input (i), and output (o) states. These gates control the flow of information through the model by using point-wise multiplications and sigmoid functions σ, which bound the information flow between zero and one by the following steps.

In the first step, the forget (f) gate controls the information that is passed through the LSTM cell. It perceive information at $h(t-1)$ and $x(t)$ and produce output numbers between 0 and 1, zero to forget and 1 to keep the information in the cell state $C(t-1)$ as seen in Figure B.5a).

$$f(t) = \sigma(W_{(x\to f)}x(t) + W(h \to f)h(t-1) + b_{(1\to f)}) \tag{B.3}$$

In the next step, input gate i with sigmoid σ layer identifies which values will be updated and with *tanh* layers creates the vector to update the state from C(t -1) to C(t).

$$i(t) = \sigma(W_{(x\to i)}x(t) + W_{(h\to i)}h(t-1) + b_{(1\to i)}) \tag{B.4}$$

$$\widetilde{C}(t) = tanh(W_{(x\to c)}x(t)) + W_{(h\to c)}h(t-1) + b_{(1\to c)}) \tag{B.5}$$

$$C(t) = f(t) * C(t-1) + i(t)\widetilde{C}(t), \tag{B.6}$$

In the last step, output is decided on the basis of the state of the cell but with filtered version. It is done by first running sigmoid σ layer that decides which information of the cell will go to the output then *tanh* evaluates the values between (-1 and 1) and multiply it with output of the input *i*gate as demonstrated in the Figure B.5d).

$$o(t) = \sigma(W_{(x\to o)}x(t) + W_{(h\to o)}h(t-1) + b_{(1\to o)}) \tag{B.7}$$

$$h(t) = o(t)tanh(C(t)), \tag{B.8}$$

where $C(t)$ is the input to the cell at time t, C is the cell, and h is the output. $W_{(x\to y)}$ are the weights from x to y.

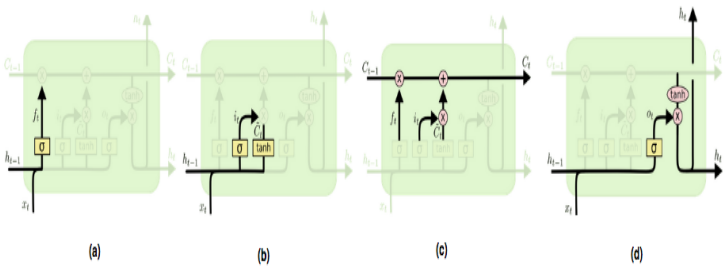


Fig. B.5: Depiction of Steps LSTM system with forget, input and output states.

In this paper, linear architecture of CNN and LSTM have been employed, to extract the facial features with the help of CNN from the input faces of TBI patients and then feed to LSTM to exploit the temporal relation on the basis of extracted features in timely manner. For feature extraction we have fine tuned the CNN with off the shelf pre-trained VGG-16CNN model [66]. Features are obtained as `fc7` layer of CNN with VGG-16 model that is feed into LSTM model to analyze the performance of combined CNN + LSTM deep neural architecture. Figure C.1 is exhibiting the main steps of this neural network along with pre-processing techniques.

3.4 Fusion of RGB and Thermal Modalities

We have employed two fusion approaches in order to analyze the performance of both RGB and thermal modalities for FER as we did it in [14]. These techniques are: a) Early fusion approach and b) Feature level late fusion approach. In the early fusion, both RGB and thermal modalities are combined into single array for feature extraction through CNN. In the feature level fusion technique, feature vectors obtained separately from RGB and thermal data with the help of CNN and then combined as a single input to LSTM model for classification. Block diagrams of both of the fusion approaches are demonstrated in Figure B.6.

4 EXPERIMENTAL RESULTS

In this section, we will discuss the database structure and its utilization in our experiments. We then demonstrate the performance of the proposed system.

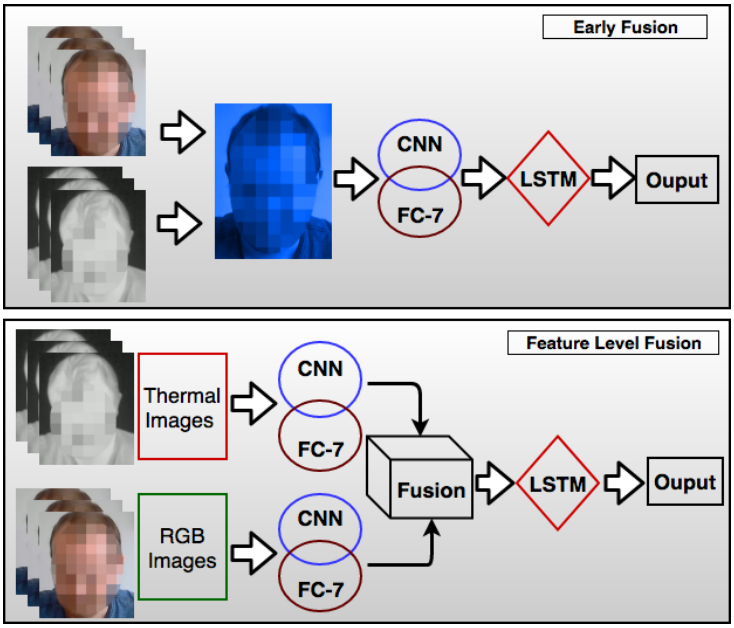


Fig. B.6: Early and Feature Level Fusion schemes for RGB and thermal modalities. Adopted from C.2 [67]

4.1 The Database Structure

TBI patients database is established for FER as described in section 3.1. We have collected data from 9 TBI patients in multiple phases in 45 sessions with the help of experts, trainers and physiotherapists in specialized scenarios. We collected the data from OSterskoven Neurocneter Hobro and SCN Frederikshavn, Denamrk. It is noteworthy that all the patients did not participate in all the activities due to nature of disability. We analyzed 6 basic emotional expression of the participated TBI patients, which are happy, sad, anger, fatigue, surprise and neutral. Data collection scenario are adjusted to get more frontal images with clearer expressions. A tablet displaying emotional scenes is placed just parallel to camera while recording their facial expressions. One interesting observation is that all the TBI patients have taken deep interest in mind game, and movie or picture illustration regardless of their disability nature. This allows us to collect more neutral, happy and angry expressions. However, we could not collect much expressions of sadness, surprised and fatigue due to non-cooperation, traumatic disabilities and other social and technical issues. Another complication is associated with large pose variation due extensive head motions of few TBI patients. We got 935 video events, each of maximum 5 seconds of duration comprising almost 140,000 frames. However, as data were highly imbalance with most of the events are captured with neutral expressions so we have applied data augmentation techniques to avoid over fitting.

4.2 Performance Evaluation

Table B.2 demonstrates the performance comparison of VJ face detector vs deep face detector. From the results it is clearly evident that deep face detector [59] is much better than VJ. This is because VJ is unable to detect faces when there is large pose variations, whereas deep face detector [59] has successfully identified the face that is 90 out of plane with 74.24% confidence. Also, VJ produce false detection of facial images due to change in illumination condition. On the other deep face detector out performed the VJ in challenging conditions with accuracy and speed. Last but not least, when face quality assessment is applied on database of VJ it produced 14,875 false frames detection that account error percentage of 12.56%. On the contrary when FQA is applied over deep face detector database, it detected 4,169 erroneous frames. That accumulates only 2.97% false frames as seen in Table B.2. These erroneous frames in both cases are either due to false detection or obstruction in front of the face resulting in lower quality of the facial images.

Table B.2: The performance comparison of Viola Jones face detector Vs Deep face detector.

	Viola Jones	**Deep Face**
Total no. of Frames	118502	140250
Training frames	94800	112200
Testing Frames	23700	28005
Miss-Detection	14875	4169
Miss Detection Percentage	12.56%	2.97%
Frames Missed	21748	0

We have employed a number of techniques to evaluate the performance of the system, such as by illustrating FER accuracy by Area Under Curve (AUC) and by displaying confusion matrix for both early and feature level fusion. Table B.3 depicts the FER accuracy by measuring AUC. In this scenario number of epochs of CNN-LSTM system is gradually increased by the steps of 5, from 5 to 50 keeping other parameters such as RHO, recurrent depth, and drop-out probability constant. From the results we observe that the accuracies of RGB database are increased with gradual increase in epochs up to 25 epochs. It reached up to level of 83.25% at the 25th epoch. On the other hand, with fusion of RGB and thermal at early stage, AUC is gradually increased to 83.54% untill 25th epochs and then decreased with further increase in epochs to 50th level. It is also observed that with thermal data only, FER accuracy is also gradually increased from 68.951 to maximum value of 73.04% at the 25th epoch. It is also noted that all the databases such RGB with non-frontalization, frontalization, thermal and fused datasets exhibited optimal performance at the 25-30 epochs.

In contrast to Table B.3, we have gradually changed the RHO values, while keeping the other parameters such as epochs, recurrent depth and drop-out probability constant as seen in Table B.4. It is observed that the RGB non- frontalized dataset reached maximum the accuracy of 79.04% when RHO value is 7. RHO value is gradually changed at steps of 2, from 1 to 13, means giving more temporal information

Table B.3: AUC results for FER of TBI patients data with gradual increase in epochs

	AUC			
Epochs	**RGB Non-Frontalized**	**RGB Frontalized**	**Thermal**	**Early Fusion RGB+T**
10	74.91	75.55	68.95	75.01
15	75.40	78.87	69.51	78.42
20	78.01	80.25	71.68	79.85
25	79.36	83.26	73.04	83.54
30	77.85	82.97	72.16	82.45
35	76.29	80.56	70.21	81.03
40	75.06	78.17	68.55	79.49
45	72.63	76.81	67.86	78.03
50	72.31	74.86	67.52	76.46

for FER. AUC values showed the RGB frontalized dataset exhibited slightly higher accuracy of 83.75% at the same 7th RHO. In contrast, thermal data got the accuracy above 70% in all steps with maximum value of 75.02% at 5th RHO and minimum value of 71.19%. Maximum AUC is observed with RGB+T early fusion data base with maximum value of 80.44% when RHO value is increased to 7. It is clearly evident from the experiment results for TBI patients data, despite of the challenging data set accuracy of system is increased to certain extent as compared to [6].

Table B.5 and TableB.6 illustrated the confusion matrix obtained by the early feature level fusion of the RGB and Thermal modalities respectively. Early fusion of both modalities has demonstrated the maximum accuracy of 88% for neutral expressions, along with 85% for angry, 82% for happy and 78% for sad emotions. However, 67% accuracy is for fatigue feelings due to the less training data for this expressions. Feature level fusion showed better results for neutral 89%, happy 85% and for fatigue 71% accuracy in TableB.6 as compared to early fusion. Both early and feature level fusion exhibited accuracy 71% for surprised feelings.

5 Conclusion

In this paper, we investigated the performance of FER for real TBI patients in uncontrolled natural challenging conditions. The study depicts the complexities that are associated with TBI patient data collection for database establishment due to varying illumination and changing pose conditions. Data is captured from TBI patients residing in neurocenters in real scenarios to have reliable data. We proposed an effective approach for FER for these subjects. Facial images are extracted from the video frames by employing different methods followed by various pre-processing techniques ensuring high quality of images that are fed into a CNN+LSTM based deep learning architecture to exploit both spatio-temporal information to detect the patients mental

Table B.4: AUC results for FER of TBI patients data with gradual increase in RHO

	AUC			
RHO	**RGB Non-Frontalized**	**RGB Frontalized**	**Thermal**	**Early Fusion RGB+T**
1	75.46	77.59	73.55	72.42
3	76.51	79.08	73.83	75.73
5	78.49	83.75	75.02	78.67
7	79.04	83.41	74.89	80.45
9	78.42	81.46	74.16	79.03
11	77.11	79.27	73.68	78.33
13	74.22	78.07	72.10	76.70

Table B.5: FER confusion matrix for early fusion of RGB and thermal modalities

	Neutral	**Happy**	**Angry**	**Sad**	**Fatigued**	**Surprised**
Neutral	**0.88**	0.03	0.02	0.04	0.02	0.01
Happy	0.04	**0.82**	0.02	0.03	0.02	0.07
Angry	0.02	0.02	**0.85**	0.05	0.06	0.02
Sad	0.06	0.01	0.04	**0.78**	0.11	0.01
Fatigued	0.07	0.01	0.05	0.2	**0.67**	0.09
Surprised	0.02	0.08	0.1	0.02	0.06	**0.71**

Table B.6: FER confusion matrix for feature level fusion of RGB and thermal modalities

	Neutral	**Happy**	**Angry**	**Sad**	**Fatigued**	**Surprised**
Neutral	**0.89**	0.02	0.03	0.05	0.01	0.01
Happy	0.03	**0.85**	0.02	0.03	0.02	0.04
Angry	0.02	0.02	**0.82**	0.05	0.06	0.02
Sad	0.05	0.01	0.04	**0.81**	0.11	0.01
Fatigued	0.06	0.01	0.05	0.3	**0.71**	0.02
Surprised	0.05	0.04	0.1	0.02	0.06	**0.71**

status in terms of facial expressions. The results are demonstrated for 6 basic facial expressions classification by using multimodal data in both early and feature level fusion. The results showed clear improvement over our previous approach in [14]. We observed that deep face detector has enhanced the detection rate of facial images even in poor lightening and extensive non-frontal images. However, for future work, TBI patients upper body movements, larger dataset for training and subject specific knowledge base incorporation can be explored for mood and emotion recognition to better facilitate rehabilitation procedure.

References

[1] D. T. Stuss and B. Levine, "Adult clinical neuropsychology: Lessons from studies of the frontal lobes," *Annual Review of Psychology*, vol. 53, no. 1, pp. 401–433, 2002, pMID: 11752491. [Online]. Available: https://doi.org/10.1146/annurev.psych.53.100901.135220

[2] H. S. Levin, D. Williams, M. J. Crofford, W. M. High, H. M. Eisenberg, E. G. Amparo, F. C. Guinto, Z. Kalisky, S. F. Handel, and A. M. Goldman, "Relationship of depth of brain lesions to consciousness and outcome after closed head injury," *Journal of Neurosurgery*, vol. 69, no. 6, pp. 861–866, 1988.

[3] I. J. B. Fary Khan and I. D. cameron, "Rehabilitation after brain injury," *The medical Journal of Australia*, vol. 178, pp. 290–295, March 2003.

[4] Brain injury facts | international brain injury association-ibia. [Online]. Available: http://www.internationalbrain.org/brain-injury-facts/

[5] T. CA, B. JM, B. MJ, and X. L., "Traumatic brain injury–related emergency department visits, hospitalizations, and deaths — united states, 2007 and 2013," *Morbidity and Mortality Weekly Report (MMWR)*, vol. 66(No. SS-9), p. 1–16, 2017.

[6] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[7] B. Dang, W. Chen, W. He, and G. Chen, "Rehabilitation treatment and progress of traumatic brain injury dysfunction," *Neural Plasticity*, vol. 2017, 2017.

[8] J. Bird and R. Parente, *Recognition of nonverbal communication of emotion after traumatic brain injury*, 2014.

[9] A. Bender, C. Adrion, L. Fischer, M. Huber, K. Jawny, A. Straube, and U. Mansmann, "Long-term rehabilitation in patients with acquired brain injury," *Deutsches Ärzteblatt International*, vol. 113, pp. 634–641, September 2016.

[10] M. P. Hyett, G. B. Parker, and A. Dhall, *The Utility of Facial Analysis Algorithms in Detecting Melancholia*. Cham: Springer International Publishing, 2016, pp. 359–375.

[11] Y. Chen, *Face Perception in Schizophrenia Spectrum Disorders: Interface Between Cognitive and Social Cognitive Functioning*. Dordrecht: Springer Netherlands, 2011, pp. 111–120.

[12] J. Klonovs, M. A. Haque, V. Krueger, K. Nasrollahi, K. Andersen-Ranberg, T. B. Moeslund, and E. G. Spaich, *Monitoring Technology*. Cham: Springer International Publishing, 2016, pp. 49–84.

[13] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial video-based detection of physical fatigue for maximal muscle activity," *IET Computer Vision*, vol. 10, no. 4, pp. 323–329, 2016.

[14] C. M. A. Ilyas, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," IEEE Signal Processing Society. IEEE Xplore, 2018.

[15] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2017.

[16] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*. Cham: Springer International Publishing, 2017, pp. 151–162.

[17] M. D. Lauterbach, P. L. Notarangelo, S. J. Nichols, K. S. Lane, and V. E. Koliatsos, "Diagnostic and treatment challenges in traumatic brain injury patients with severe neuropsychiatric symptoms: insights into psychiatric practice," *Neuropsychiatr Dis Treat*, vol. 11, pp. 1601–1607, 2015.

[18] K. Lander and S. Metcalfe, "The influence of positive and negative facial expressions on face familiarity," *Memory*, vol. 15, no. 1, pp. 63–69, 2007, pMID: 17479925. [Online]. Available: https://doi.org/10.1080/09658210601108732

[19] Y. I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, Feb 2001.

[20] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, April 2006.

[21] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, Feb 2014.

[22] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[23] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Constructing facial expression log from video sequences using face quality assessment," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, Jan 2014, pp. 517–525.

[24] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. N. Yanushkevich, "Gauss–laguerre wavelet textural feature fusion with geometrical information for facial

expression identification," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, p. 17, Sep 2012.

[25] R. N. Anwar Saeed, Ayoub Al-Hamadi and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in Human-Computer Interaction*, vol. 2014, pp. 1–13, 2014.

[26] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[27] Y. li Tian, "Evaluation of face resolution for expression analysis," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, June 2004, pp. 82–82.

[28] H. Li and G. Hua, "Hierarchical-pep model for real-world face recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4055–4064.

[29] Z. Huang, R. Wang, S. Shan, and X. Chen, "Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning," *Pattern Recognition*, vol. 48, no. 10, pp. 3113 – 3124, 2015, discriminative Feature Learning from Big Data for Visual Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320315001120

[30] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *CoRR*, vol. abs/1603.05474, 2016. [Online]. Available: http://arxiv.org/abs/1603.05474

[31] Y. Tang, "Deep learning using support vector machines," *CoRR*, vol. abs/1306.0239, 2013. [Online]. Available: http://arxiv.org/abs/1306.0239

[32] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550. [Online]. Available: http://doi.acm.org/10.1145/2522848.2531745

[33] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1805–1812.

[34] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 435–442. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830595

[35] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Computer Vision – ACCV 2014*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, pp. 143–157.

[36] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun 2016. [Online]. Available: https://doi.org/10.1007/s12193-015-0209-0

[37] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baró, S. J. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of deceptive facial expressions of emotion," *CoRR*, vol. abs/1707.04061, 2017. [Online]. Available: http://arxiv.org/abs/1707.04061

[38] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 494–501. [Online]. Available: http://doi.acm.org/10.1145/2663204.2666274

[39] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI 2016. New York, NY, USA: ACM, 2016, pp. 445–450. [Online]. Available: http://doi.acm.org/10.1145/2993148.2997632

[40] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[41] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[42] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[43] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," *Sensors*, vol. 11, no. 10, pp. 9573–9588, 2011.

[44] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using hog–ebgm," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537 – 1543, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508001104

[45] S. Berretti, B. Ben Amor, M. Daoudi, and A. del Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, p. 1021, Jun 2011. [Online]. Available: https://doi.org/10.1007/s00371-011-0611-x

[46] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803 – 816, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608001844

[47] M. Z. Uddin and M. M. Hassan, "A depth video-based facial expression recognition system using radon transform, generalized discriminant analysis, and hidden markov model," *Multimedia Tools and Applications*, vol. 74, no. 11, pp. 3675–3690, Jun 2015.

[48] G.-J. de Vries, S. Pauws, and M. Biehl, *Facial Expression Recognition Using Learning Vector Quantization*. Cham: Springer International Publishing, 2015, pp. 760–771.

[49] A. Ravichander, S. Vijay, V. Ramaseshan, and S. Natarajan, *Automated Human Facial Expression Recognition Using Extreme Learning Machines*. Cham: Springer International Publishing, 2016, pp. 209–222.

[50] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ser. ICMR '15. ACM, 2015, pp. 643–650.

[51] D. Triantafyllidou and A. Tefas, "Face detection based on deep convolutional neural networks exploiting incremental facial part learning," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 3560–3565.

[52] H. Yoshihara, M. Seo, T. H. Ngo, N. Matsushiro, and Y. W. Chen, "Automatic feature point detection using deep convolutional networks for quantitative evaluation of facial paralysis," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Oct 2016, pp. 811–814.

[53] R. Kharghanian, A. Peiravi, and F. Moradi, "Pain detection from facial images using unsupervised feature learning approach," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 419–422.

[54] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Quality-aware estimation of facial landmarks in video sequences," in *2015 IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 678–685.

[55] A. Barman, A. Chatterjee, and R. Bhide, "Cognitive impairment and rehabilitation strategies after traumatic brain injury," *Indian Journal of Psychological Medicine*, vol. 38, no. 3, pp. 172–181, May-Jun 2016.

[56] K. McKenna, D. M. Cooke, J. Fleming, A. Jefferson, and S. Ogden, "The incidence of visual perceptual impairment in patients with severe traumatic brain injury," *Brain Injury*, vol. 20, no. 5, pp. 507–518, 2006. [Online]. Available: https://doi.org/10.1080/02699050600664368

[57] T. Tsaousides and W. A. Gordon, "Cognitive rehabilitation following traumatic brain injury: assessment to treatment," *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 76, no. 2, pp. 173–181, 2009. [Online]. Available: http://dx.doi.org/10.1002/msj.20099

[58] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[59] Face detection with opencv and deep learning. [Online]. Available: https://www.pyimagesearch.com/2018/02/26/face-detection-with-opencv-and-deep-learning/

[60] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 532–539.

[61] R. Irani, K. Nasrollahi, A. Dhall, T. B. Moeslund, and T. Gedeon, "Thermal super-pixels for bimodal stress recognition," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.

[62] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4295–4304.

[63] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–1, 2017.

[64] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[65] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 369–389.

[66] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[67] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

Paper C

Rehabilitation of Traumatic Brain Injured Patients: Patient Mood Analysis from Multimodal Video

Chaudhary Muhammad Aqdus Ilyas, Kamal Nasrollahi,
Matthias Rehm and Thomas B. Moeslund

The paper has been published in the
Proceedings of 25th IEEE conference on Image Processing (ICIP 2018)
pp. 2291-2295, 2018. doi: 10.1109/ICIP.2018.8451223

The layout has been revised.

Abstract

Rehabilitation after traumatic brain injury (TBI) is very critical as it is largely unpredictable depending upon the nature of the injury. Rehabilitation process and recovery time also varies, as it takes months and years, depending upon the assessment of treatment, mental and physical conditions and strategies. Due to non-cooperative behaviour of patients, and increase in negative emotional expressions it is very beneficial to evaluate these expressions in a contactless way, and perform a rehabilitation physiotherapy, cognitive or other behavioral activities when the patient is in a positive mood. In this paper we have analyzed the methods for facial features extraction for TBI patients to determine optimal time to have aforementioned rehabilitation process on the basis of positive and negative facial expressions. We have employed a deep learning architecture based on convolutional neural network and long short term memory on RGB and thermal data that were collected in challenging scenarios from real patients. It automatically identifies the patient's facial expressions, and inform experts or trainers that "it is the time" to start rehabilitation session.

1 Introduction

Traumatic brain injury (TBI) causes life-long damage to cognitive, physical, behavioural and social functions. It may take up to 5 years or more for recovery after TBI [1]. According to International Brain Injury Association (IBIA), annually one million people suffer from traumatic brain injury (TBI) only in America whereas same number of people suffer with TBI in Europe [2]. American Center for Disease Control and Prevention estimates more than 3.7 million people are living with long term disability after TBI. During rehabilitation period, patient has to live in a specialized care center called neuro-center or care home where the main focus is on the retraining of activities of daily life, cognitive, social and physical exercises through a set of protocols. Recovery targets are based on determination of combination of cognitive, behavioral and physical shortfalls. It is seen that rehabilitation activities are performed daily on set time table of neuro-center, regardless of mental conditions of subject. This leads to more time expensive training with less result oriented outcome.

There is high urgency of fast and accurate rehabilitation process so the TBI patients have to spend less time in care centers or have to suffer less with limited independence and low quality of life. Caregivers, trainers or experts dealing with TBI patients face severe difficulty in performing rehabilitation activities as the patients have limited or reduced ability to perceive social and interaction signals [3]. In addition to that there is relative increase in negative emotions like depression, anger, anxiety, sadness, verbal or physical aggression and lack of social communication after TBI [4] [5]. Extra consideration and care need to be made while interacting with these patients. Experts and trainers believe that with assessment of impact of injury to positive and negative emotions, caregivers can provide more accurate and faster rehabilitation services [6]. Goal and activity setting, for brain injury rehabilitation by involving patients emotional states, increase the chances of faster recovery with broader aspects [5]. It will provide flexibility to staff to work around with many more patients at the same neuro-center in less time.

Experts are putting emphasis on implementing Computer Vision (CV) techniques in health care sector as population is growing, so as the number of brain injured patients. Therefore, automatic diagnosis of mental and physical health states through unobtrusive computer vision techniques by using facial features has rapidly increased since past decades [7] [8] [9]. The fundamental approach for utilizing these CV techniques is to diminish the errors by human assessment. Furthermore, these approaches are cost effective as compared to medical examination by physicians or doctors, and can provide continues monitoring of the patients.

Existing CV techniques for facial expression recognition (FER) systems are mostly designed and implemented for healthy people. However, TBI patients' emotional states are quite different from healthy people as they have high degree of imbalance of six common emotional expressions accompanied by reduced muscle movement or paralysis. The database established for TBI patients for FER described in our previous paper [4], shows that it is very difficult to have all six expressions. Therefore, in this paper we suggest to classify the facial features into two emotional states either positive and negative. If patients are found to be in a positive mood, the caregivers are alarmed to start the rehabilitation. Furthermore, we do bimodal analysis of facial images in both the color RGB and thermal modalities. To do this, we have expanded our previous database of [4] by including more TBI patients. Experts and psychologists have been asked to help us annotating the collected data. They characterized positive expression as smile, laugh, surprise and few unique neutral expressions, while fear, disgust, anger, sad, stress and fatigue are categorized as negative expressions, sometimes additionally associated with lips trembling, teeth grinding and frequent eye blinking [10]. In case of TBI patients, negative expressions are more frequent as compared to positive ones. Our obtained experimental results using deep learning techniques show that the two employed modalities can complement each other on classifying patients status to positive or negative.

In terms of methodology, contributions by are probably most close to our method but these systems work well for healthy people in controlled environment. Moreover these systems have luxury of data sets where subjects were cooperative with no or less pose variation, minimum occlusions and high quality images unlike with TBI patients. As described in our previous paper [4], our database in [4] was established with Face Quality Assessment (FQA) but with only contained RGB images. In the current paper, we have improved the database with both RGB and thermal images with additional subjects and more pre-processing techniques like face frontalisation. We have verified the proposed system with real data of TBI patients collected in real environment at neuro-center where these TBI individuals are looked after 24/7.

The rest of this paper is organized as follows: The related work on FER are reviewed in the next section. Section 3 describes the new database including data collection and pre-processing techniques. Section 4 describes the proposed methodology for facial feature extraction and expression recognition. Section 5 presents the results obtained from the experiments. Finally, Section 7 concludes the paper.

2 Related Work

Current FER system can be categorized on the basis of methods used for feature extraction and classification. Our main focus is on the methods involving Convolution Neural Networks (CNN) or other deep learning approaches as they provide state of the art results for, e.g., face recognition [11] [12] [13], facial expressions recognition [14] [15] [16] [17] [18] [19] [20] [21] [22] and emotional states identification [23] [24] [25] [26]. Handcrafted features such as Local Binary Pattern (LBP), SIFT, Local Quantized Pattern (LPQ) and Histogram of Oriented Gradients (HOG) applied in [27] [28] [29] [30] [31] are outperformed by CNN based deep neural networks despite their low computational cost.

In [14], Tang proposed deep CNN along with Support Vector Machines (SVM) and achieved state of the arts results for FER with 1st prize in FER-2013 competition. In 2014, Liu [16] performed three functions- feature learning, feature selection and classification in unified manner through Boosted Deep Belief Networks (BDBN). This method worked exceptionally well even for extremely complicated features from facial image. [19] used DBN models to overcome the limitations of linear feature selections. Yu and Zang [17] in 2015, presented their work for Emotion recognition in Wild challenge for image based static FER. They have applied multiple deep CNN with random initialization of each network and minimized likelihood and hinge loss. Their results surpassed the challenge baseline significantly. In year 2017, [22] exercised CNN to learn features from VGG-Faces and integrated with Long Short Term Memory (LSTM) to gain the temporal information. This approach was further improved by [21] who applied deep CNN for features classification into expressions and feed the system with super-resolved facial images.

3 TBI Patient Database for FER

3.1 Data Acquisition

To analyze facial expressions, data is collected in three pre-specified scenarios from seven TBI patients in two modalities: RGB and Thermal. Pre-specified scenarios in data collection are maintained to have reliable data for further use. Those scenarios are: 1) cognitive activity 2) physiotherapy and 3) social communication. These scenarios are selected after consulting many experts and care givers, who are working on rehabilitation of TBI individuals in Denmark. On contrary to healthy people, as mentioned in [4], data acquisition task is quite complicated due to extreme behavioural responses, verbalization, physical aggression, impaired reasoning, reduced cognitive skills along with frequent pose variations.

Ilyas et al. [4], collected RGB database by Axis RGB-Q16 camera with resolution of 1280 x 960 to 160 x 90 pixels at 30fps (frames per second) and applied pre processing techniques of face detection, FQA, (Supervised Decent Method) SDM for landmark detection and tracking before logging into a face log. We have operated with a Logitech camera as well to record the starting and ending time stamp of particular expressions. Along with RGB, we have gathered thermal images of TBI subjects

with Axis Thermal-Q1922 camera with focal lens of 10 mm. RGB cameras are prone to difficulties in challenging conditions like shadows or when subject are obscured with complex background. Thermal cameras, on the other hand, can provide addition information of a scene. Thermal and RGB imagery are synchronized with the help of time stamps and annotation are made in sequence of facial expressions. Both RGB and thermal images are collected with same 30 fps. Furthermore, homography estimation is employed for image registration by determining homography matrices from RGB to thermal by [32].

Table C.1: Database Of TBI patients with Activity Participation

Subjects	Number of Sessions	Activities Participated		
		Cognitive	Social Comm	Physiotherapy
Subject A	7	Y	Y	Y
Subject B	5	Y	Y	Y
Subject C	5	Y	Y	Y
Subject D	7	Y	X	Y
Subject E	3	Y	X	X
Subject F	4	Y	Y	Y
Subject G	3	X	Y	Y

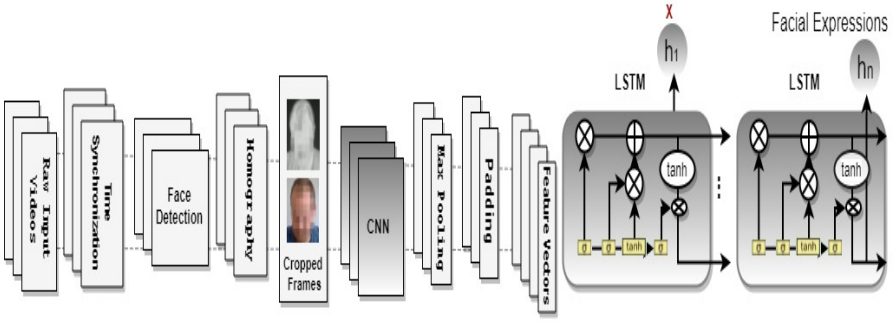


Fig. C.1: CNN+LSTM based deep learning architecture for both modalities to exploit spatio-temporal information for FER.

3.2 Database Structure

Data is collected from seven TBI patients in 34 sessions on the above mentioned three pre-specified scenarios. Few subjects did not take part in all activities, details are described in Table C.1. Two categories of expressions are recorded: Positive Expression (PE) and Negative Expressions (NE). PEs are smile, laugh, surprise and few unique neutral expressions, while NEs are fear, disgust, anger, sad, stress and fatigue. We have got 861 video events, each of maximum 5 seconds in length.

4 The Proposed Methodology

This section presents the architecture of the intended approach for FER analysis of real TBI patient in realistic environment. We have employed the same method as followed in [4] but employed new pre-processing technique of face frontalization because of large pose variation. We tested the deep learning method of [4] on both modalities, with early and late fusions. Facial expressions are recognized by employing CNN (to use spatial features) and linking with LSTM to utilize spatio-temporal attributes of RGB, thermal and fused RGB-thermal modalities. The block diagram of the proposed method is illustrated in Figure C.1. The steps of the proposed system are further explained in the following subsections.

4.1 Pre-Processing

Firstly, the face is detected, and facial landmarks are identified and tracked using [33] from a synchronized input video. TBI patients have large pose variations so to avoid loss of information, the posed faces are rotated using a frontalization algorithm. For face frontalization, landmarks are calculated with arbitrary facial positions and by finding inverse of the transpose matrix, the face is frontalized. In next step, face cropping is done in RGB modality, and associated faces in thermal modality is cropped by applying a homography. Homography is a special technique that allows geometric transformation of fixed points from one plane to another. In this case, RGB and thermal planes are homo-graphed with subject face. To remove erroneous detection and ensuring high quality of images, face quality assessment is applied before feeding the faces into the CNN pipeline.

4.2 CNN + LSTM Architecture

After the pre-processing of the data, it is fed to 2D-CNN for training purpose for mood recognition based on PE and NE. This network is fine tuned by VGG-16 face model [34] for spatial feature extraction. CNN parameters are initialized randomly and through back propagation using gradient descent its weights are adjusted. Thermal data is also fine tuned with pre-trained VGG-16 face (RGB) model. CNN deals with frames in isolated manner. For capitalizing on relation with time, special Recurrent Neural Network (RNN) called LSTM is employed. LSTM is gate controlled network with input (i), output(o) and forget (f) gates. LSTM gates holds the input information as long as its forget gate is not triggered to acquire the temporal information between frames for said purposes. These gates control the flow of instructions by point wise multiplication and sigmoid functions σ, which bound the information flow between zero and one by the followings:

$$i(t) = \sigma(W_{(x\to i)}x(t) + W_{(h\to i)}h(t-1) + b_{(1\to i)}) \tag{C.1}$$

$$f(t) = \sigma(W_{(x\to f)}x(t) + W(h\to f)h(t-1) + b_{(1\to f)}) \tag{C.2}$$

In these equations, W are weights associated with activated neurons for particular input i. Where as σ squashes the value of activation between the range of 0 and 1

$$z(t) = tanh(W_{(x\to c)}x(t)) + W_{(h\to c)}h(t-1) + b_{(1\to c)}) \tag{C.3}$$

$$c(t) = f(t)c(t-1) + i(t)z(t), \tag{C.4}$$

$$o(t) = \sigma(W_{(x\to o)}x(t) + W_{(h\to o)}h(t-1) + b_{(1\to o)}) \tag{C.5}$$

$$h(t) = o(t)tanh(c(t)), \tag{C.6}$$

where $z(t)$ is the input to the cell at time t, c is the cell, and h is the output. $W_{(x\to y)}$ are the weights from x to y. In the classification, LSTM finally provides a decision score for the expression recognition.

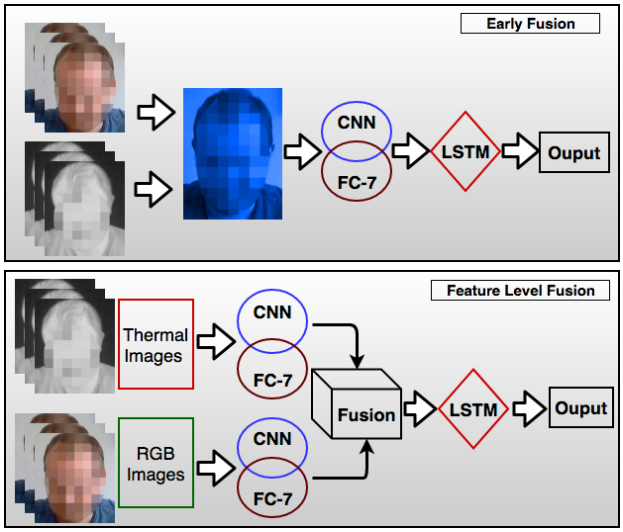


Fig. C.2: Block diagram of early and Feature Level Fusion of modalities for FER.

4.3 Fusion Scheme

In order to analyze the ability of both modalities in FER applications, two approaches were employed: 1) data level fusion (early) 2) feature level Fusion. In the first approach both modalities are combined into data array for feature learning through CNN. In the second method, both RGB and thermal imagery features are fed separately into deep learning system for feature learning and combined together as input for second classifier (LSTM) for final output. Block diagram of both modalities can be seen in Figure C.2.

Table C.2: Confusion matrix by feature level fusion of modalities for 6 basic FER.

	Neutral	Happy	Angry	Sad	Fatigued	Surprised
Neutral	0.77	0.03	0.02	0.07	0.07	0.01
Happy	0.04	0.71	0.02	0.03	0.05	0.16
Angry	0.04	0.02	0.81	0.09	0.03	0.02
Sad	0.07	0.01	0.05	0.76	0.13	0.01
Fatigued	0.09	0.01	0.09	0.1	0.55	0.11
Surprised	0.07	0.14	0.1	0.02	0.06	0.56

5 Experimental Results

We demonstrate the results in the following contexts:

a) Classification of six basic expression groups in both early and feature level fusion scenarios to evaluate the performance of CNN+LSTM based FER

b) PE and NE classifications before and after face frontalization on all individual modalities and fusions.

First we produced results of positive and negative mood identification (based on PE and NE) without employing face frontalization (FF) and then with face frontalization. It is seen in table 4 column 1-4, after FF recognition accuracy is increased to 86.93 percentage from 79.34. In second case, we trained our system for thermal data, true positive and true negative are 69 and 65 percentage with high miss classification rate of 23.74 percentage. Overall recognition accuracy is achieved up to 74.45 percentage. In next stage we combined both RGB with FF to thermal data in early fusion scheme and obtained accuracy of 84.39 percentage for mood recognition. We also employed early and feature level fusion to analyze the results for 6 common facial expressions in Table B.5 and Table C.2. In both cases, fatigue and surprise have less recognition accuracy due to less available data. If we compare table 4 with table B.5 and C.2, we can see that accuracy of system is increased for positive and negative expressions as compared to all 6 expressions. In the next stage we employed the [22] system on our database 4. It is observed that its accuracy is 87.97 percentage much lesser than [22] 97.2 percentage, when he implemented on CK+ database. In last stage, we employed the feature level fusion and achieved 89.74 percentage of accuracy. By feature level fusion, despite computational expensive surpassed other state of art methods for positive and negative expression recognition. That shows that our system is producing competitive results with challenging data sets.

6 Conclusions

Mood recognition is important task for rehabilitation and care centers. In this work we have faced the challenge of mood recognition of TBI patients rather than facial expression recognition for healthy people. In case of TBI individuals, extraction of all expression is very complicated and its dependant to patient disability and FER

Table C.3: Recognition accuracy of proposed method in different contexts

Confusion Matrix %	RGB Non-Frontal		RGB Frontal		Thermal		Early Fusion		[22]		Feature Level Fusion	
	PE	NE	PE	NE	PE	NE	PE	NE	PE	NE	PE	NE
Positive Expression (PE)	0.75	0.17	0.86	0.15	0.69	0.25	0.84	0.14	0.79	0.12	0.86	0.11
Negative Expression (NE)	0.21	0.71	0.11	0.87	0.21	0.65	0.16	0.79	0.1	0.82	0.09	0.89
Recognition Accuracy (%)	79.34		86.93		74.45		84.39		87.97		89.74	

Table C.4: Confusion matrix by early fusion of modalities for 6 basic FER

	Neutral	Happy	Angry	Sad	Fatigued	Surprised
Neutral	0.77	0.03	0.02	0.07	0.07	0.01
Happy	0.04	0.71	0.02	0.03	0.05	0.16
Angry	0.04	0.02	0.81	0.09	0.03	0.02
Sad	0.07	0.01	0.05	0.76	0.13	0.01
Fatigued	0.09	0.01	0.09	0.1	0.55	0.11
Surprised	0.07	0.14	0.1	0.02	0.06	0.56

did not provide good results [4]. However, we recognized the mood of patients with accuracy of 86.93 percentage that is very close to [22] system when implemented on TBI patient database. So this system can help physiotherapist and trainers in fast rehabilitation process after recognizing the positive mood of the patient. Furthermore, we applied early and feature level fusion to enhance the recognition rate of the system. Our system results can be improved further by employing 3D face frontalization. Even though the results are encouraging, efforts are still in progress to provide the robust solutions to deal with real time and environment challenges like real time computation or patient positioning.

References

[1] I. J. B. Fary Khan and I. D. cameron, "Rehabilitation after brain injury," *The medical Journal of Australia*, vol. 178, pp. 290–295, March 2003.

[2] Brain injury facts | international brain injury association-ibia. [Online]. Available: http://www.internationalbrain.org/brain-injury-facts/

[3] J. Bird and R. Parente, *Recognition of nonverbal communication of emotion after traumatic brain injury*, 2014.

[4] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[5] B. Dang, W. Chen, W. He, and G. Chen, "Rehabilitation treatment and progress of traumatic brain injury dysfunction," *Neural Plasticity*, vol. 2017, 2017.

[6] A. Bender, C. Adrion, L. Fischer, M. Huber, K. Jawny, A. Straube, and U. Mansmann, "Long-term rehabilitation in patients with acquired brain injury," *Deutsches Ärzteblatt International*, vol. 113, pp. 634–641, September 2016.

[7] F. Li, C. Zhao, Z. Xia, Y. Wang, X. Zhou, and G.-Z. Li, "Computer-assisted lip diagnosis on traditional chinese medicine using multi-class support vector machines," *BMC Complementary and Alternative Medicine*, vol. 12, no. 1, p. 127, Aug 2012.

[8] M. P. Hyett, G. B. Parker, and A. Dhall, *The Utility of Facial Analysis Algorithms in Detecting Melancholia*. Cham: Springer International Publishing, 2016, pp. 359–375.

[9] Y. Chen, *Face Perception in Schizophrenia Spectrum Disorders: Interface Between Cognitive and Social Cognitive Functioning*. Dordrecht: Springer Netherlands, 2011, pp. 111–120.

[10] K. Lander and S. Metcalfe, "The influence of positive and negative facial expressions on face familiarity," *Memory*, vol. 15, no. 1, pp. 63–69, 2007, pMID: 17479925. [Online]. Available: https://doi.org/10.1080/09658210601108732

[11] H. Li and G. Hua, "Hierarchical-pep model for real-world face recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4055–4064.

[12] Z. Huang, R. Wang, S. Shan, and X. Chen, "Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning," *Pattern Recognition*, vol. 48, no. 10, pp. 3113 – 3124, 2015, discriminative Feature Learning from Big Data for Visual Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320315001120

[13] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *CoRR*, vol. abs/1603.05474, 2016. [Online]. Available: http://arxiv.org/abs/1603.05474

[14] Y. Tang, "Deep learning using support vector machines," *CoRR*, vol. abs/1306.0239, 2013. [Online]. Available: http://arxiv.org/abs/1306.0239

[15] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550. [Online]. Available: http://doi.acm.org/10.1145/2522848.2531745

[16] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1805–1812.

[17] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 435–442. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830595

[18] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Computer Vision – ACCV 2014*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, pp. 143–157.

[19] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun 2016. [Online]. Available: https://doi.org/10.1007/s12193-015-0209-0

[20] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baró, S. J. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of deceptive facial expressions of emotion," *CoRR*, vol. abs/1707.04061, 2017. [Online]. Available: http://arxiv.org/abs/1707.04061

[21] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*. Cham: Springer International Publishing, 2017, pp. 151–162.

[22] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2017.

[23] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 494–501. [Online]. Available: http://doi.acm.org/10.1145/2663204.2666274

[24] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI 2016. New York, NY, USA: ACM, 2016, pp. 445–450. [Online]. Available: http://doi.acm.org/10.1145/2993148.2997632

[25] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[26] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[27] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[28] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," *Sensors*, vol. 11, no. 10, pp. 9573–9588, 2011.

[29] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using hog–ebgm," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537 – 1543, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508001104

[30] S. Berretti, B. Ben Amor, M. Daoudi, and A. del Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, p. 1021, Jun 2011. [Online]. Available: https://doi.org/10.1007/s00371-011-0611-x

[31] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803 – 816, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608001844

[32] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[33] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 532–539.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

Paper D

Deep Emotion Recognition through Upper Body Movements and Facial Expression

Chaudhary Muhammad Aqdus Ilyas, Rita Nunes, Kamal Nasrollahi, Matthias Rehm and Thomas B. Moeslund

The paper has been accepted in the
Proceedings of 16th International Conference on Computer Vision Theory and Applications (VISAPP 2021) 2021.
Please note that above article is updated slighlty. The updates are:
An additional reference has been included

21 citations to this reference are added. The additional reference is a Master's thesis: Rita Nunes, *Deep Emotion Recognition through Upper Body Movements and Facial Expression*, Master thesis, Aalborg University, Denmark, 2018.
There is an overlap between the contents of that thesis and content of the article. The citation are added to make this overlap evident.

The layout has been revised.

Abstract

Despite recent significant advancements in the field of human emotion recognition, applying upper body movements along with facial expressions present severe challenges in the field of human-robot interaction. This article presents a model that learns emotions through upper body movements and corresponds with facial expressions. Once this correspondence is mapped, tasks such as emotion and gesture recognition can easily be identified using facial features and movement vectors. Our method uses a deep convolution neural network trained on benchmark datasets exhibiting various emotions and corresponding body movements. Features obtained through facial movements and body motion are fused to get emotion recognition performance. We have implemented various fusion methodologies to integrate multimodal features for non-verbal emotion identification. Our system achieves 76.8% accuracy of emotion recognition through upper body movements only, surpassing 73.1% on the FABO dataset. In addition, employing multimodal compact bilinear pooling with temporal information surpassed the state-of-the-art method with an accuracy of 94.41% on the FABO dataset. This system can lead to better human-machine interaction by enabling robots to recognize emotions and body actions and react according to their emotions, thus enriching the user experience.

1 Introduction

Human emotions play a vital role in human-human and human-machine interaction. Emotions represent the instantaneous mental states, which varies according to human behavior and communication. Researchers are emphasizing automatic recognition of human emotions as it is one of the essential parameters for natural human-machine interaction.

In human-machine interaction, the interaction would be impaired if machines cannot recognize or understand human emotions. Similar applies to human-human interaction if the other party fails to understand these body expressions.

If machines can react to our moods, that would enable smart homes or centers to adjust lighting, music, and temperature accordingly. It would also help medical doctors and physiologists automatically identify the symptoms of hypertension, depression, and other behavioral disorders, enabling them to have early preparations for such conditions. This skill can enable sociable robotics to assist people in simple tasks such as delivering meals or vacuuming the house. Humanoid robots that provide services to people, the human-robot interaction would greatly improve if these robots could adjust their reactions to the current emotional state of a person [1–4]. Generally, it would enable machines to respond, not limited to direct commands but with the ability to adjust their reactions to have natural and human-like, human-machine interaction. However, these interactions are minimal and could be improved if the robot had more knowledge about the person they need to interact with [5].

Human recognize and demonstrate emotions through multi-modalities such as through facial expressions [6–9], body movements [6, 7, 10, 11], speech recognition [12] and physiological signals [13–17]. Existing methods for identifying these body expressions are heavily relying on audio-visual cues [12] and wearable sensors such as ECG monitors [13–17]. Audio-visual fusion have achieved remarkable results with accu-

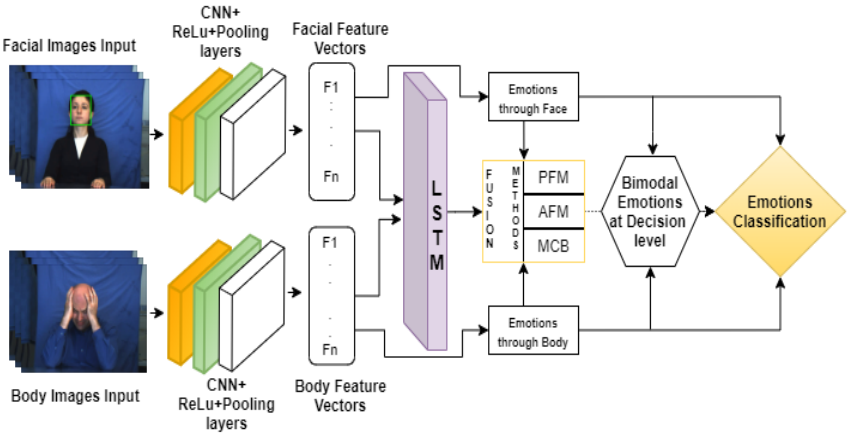


Fig. D.1: Bimodal Emotion Classification Model through Facial Expressions and Upper Body Movements.

racy of approximately 99% [18]. However, these approaches have their limitations as audio-visual sensors cannot extract inner affection [17, 19]. For instance, a person can be happy or sad without smiling and crying and vise versa. In addition, people vary greatly in the expression of their emotions. Detection of signals through physiological sensors correlate heartbeat, blood pressure, and others signals with happiness, anger, surprise, and others. This approach is more suitable for identifying inner feelings as it provides information about heart rhythm interaction with the brain system. However, the wearable body sensors cause inconvenience, so it is not suitable for emotion detection in everyday practices.

Research has also shown that body language comprises a significant amount of the affective information [10, 11]. According to Mehrabian [20], only 7% of human communication is conveyed through words, 38% through vocal tone, and 55% through non-verbal elements such as facial expression, body language, and gestures. Body posture, gestures, eye movement, hand and head movement, touch, or even personal space represent the body language [20]. Many studies have proved theoretically and empirically the benefit of incorporating various modalities in the perception of human emotions compared to using a single method [17, 21]. Complex human emotions can be fully-implied by integrating significant features from multiple modalities (e.g., facial features and body gestures).

In our research article, we have tried to explore the effectiveness of facial expressions and body gestures to recognize emotions. For this purpose, we have followed two approaches; in the first technique, Convolutional Neural Networks (CNNs) classify emotions without considering temporal features. In this system, single images are used, thus classifying each frame in videos in real-time. In the second approach, following [7, 22], we have used the temporal information to classify the emotions, but a contrast to [7, 22] full images are fed to the Long Short Term Memory (LSTM) network to exploit the temporal information. Besides, this article will explore various fusion techniques like product fusion method (PFM), average fusion method (AFM),

and Multimodal Compact Bilinear Pooling (MCB) [23] fusion and discuss their performances. Even though deep learning approaches show improved accuracy compared to conventional approaches, they are still more computationally demanding. Hence, this work will explore a solution with less computational requirements.

The paper organization is as follows: The following section 2 will discuss the related research. Section 3 presents the proposed model, illustrating how it deals with a frame-based and sequence-based recognition of the emotions with different fusion techniques. This section also discusses the evaluation of the parameters to decrease the computation power. Section 4 describes the experimental results and compares them with state-of-the-art methods. The conclusion and discussion of the experimental results with future work are presented in section 5.

2 Related Research

Emotion recognition through non-verbal modalities like facial expressions and body gestures is viewed as one of the most effective cues [19]. Therefore, many researchers have explored the fusion of visual-modalities for improved affect understanding [13, 17, 24–27]. They illustrate that facial expressions and body gestures augment each other in understanding emotional states in activities of daily living (ADL) and social robot interactions. Researchers [28] and [26] have analyzed the facial features with body gestures, particularly upper limbs and head movements, for emotion recognition. Former has utilized Former has utilized facial action units (AU) and performed classification with Bayes Net with early and late fusion whereas later has employed Spatio-temporal features classification with SVM along with Canonical Correlation Analysis (CCA) at the decision level facial action units (AU) and performed classification with Bayes Net with early and late fusion whereas later has employed Spatio-temporal features classification with SVM along with Canonical Correlation Analysis (CCA) at the decision level. In recent years, researchers have focused on deep learning approaches to solve this issue, which has achieved the best recognition rates.

Gunes and Piccardi [29] presents the performance of facial expressions, body movements, and their fused representation for automatic recognition of emotions. They extract the facial and bodily features separately and compare their performance accuracy. For facial expressions, they localize face and track landmarks and then extract features. Similarly, for body motion analysis, they track hand, shoulder, and head movements and then extract a series of features with different features representation techniques. Both facial and bodily features are classified with Support Vector Machines (SVM) and Random Forests. In the last step, features are fused to recognize emotions. Studies exhibit better performance of system with feature fusion techniques [25] [29].

"Recent work of [7] and [22] used Convolutional Neural Networks (CNN) to recognize emotion from both face and body movements [4]. Both studies incorporated temporal features into their classification, which forces them to analyze an entire video before it can be classified" [4].

2.1 Facial Expression Recognition

Facial expressions are one of the vital source to know about mood and feelings in an interpersonal communication. Therefore, researchers focus to analyze facial expressions through traditional machine leaning and advanced deep learning algorithms. [30] used a deep learning approach to merge CNNs with RNNs and evaluate how each portion of the neural network contributed to the overall success of the emotion recognition system. For training, two different architectures were used, the first with a single CNN frame and the second with a combination of CNN and RNN. Although CNN learns valuable features from the video data from the single frame regression, it disregards temporal information. This knowledge can be implemented through the use of RNN. *"Results determined that the CNN+RNN model translates to more accurate predictions"* [4].

2.2 Emotion Recognition through Body Movements

"Upper body movement, such as hand and head movement, conveys vital information related to emotional states. For instance, when a person displays a neutral emotion, they generally do not move their arms; however, when they are happy or sad, the body tends to be extended, and the hands move upwards closer to the head" [4, 18]. However, this information is subjective, dependant upon the personal attitude to the circumstances and cultural bias.Research presented by [31] suggest real-time emotion recognition through body movements and gestures. *"The features are extracted from 3D motion clips containing full-body movements, recorded using two systems: a professional optical motion capture system and Microsoft Kinect. The body joints are tracked, and feature vectors of the movements are extracted and used for classification using a linear SVM classifier. The emotions tested were the six standard emotions. Human validation demonstrated that three emotions were easily recognized from body movements (happiness, sadness, and anger), while the others (surprise, disgust, and fear) were confused with each other. Because of that, a sub-problem with only four emotions (happiness, sadness, anger, and fear) was formulated. The approach showed better results when only classifying four out of the six emotions"* [4].

" [32] took a different approach, which analyzed affective behavior solely based on upper-body movements. A range of twelve different emotions was classified according to their valence and arousal. Features were extracted from two videos, one that displayed a frontal view of the subjects and another that displayed a lateral view. The trajectories of the head and hands were tracked, and low-level physical measures, i.e., position, speed, acceleration, were extracted. Higher-level expressive and dynamic features include smoothness and continuity of movement, spatial symmetry of the hands, gesture duration were then computed, forming a 25-features vector" [4]. PCA was later applied to reduce the dimensionality of the data. Furthermore, clustering was used to classify the data into four clusters according to the categorical variables, i.e., valence (positive, negative) and arousal (high, low). *"The framework was tested on the GEMEP (GEneva Multimodal Emotion Portrayals) dataset [33]. The results demonstrate that gestures can be effectively used to detect human emotional expression"* [4].

Research proposed by [7] model these upper body movements for emotional classification by using FABO [24] dataset. They present the motion by an additional layer on the network that tracks the frame-wise difference in each sequence. This repre-

sentation involves the structure and information of the gesture/motion with the aid of weighted shadows. Another study by the same authors [34] extracts spatial and temporal features of gesture sequences through Deep Neural Networks (DNN) to generate a motion representation. *" Moreover, a Multichannel Convolutional Neural Network (MCCNN) is used to learn and extract features from the previously generated motion representation and uses such features to classify different gestures"* [4].

We have trained our system with full frames of the FABO dataset with facial and body gestures features to capture the gesture information. The networks extract the spatial and temporal information using CNN and LSTM and finally classify the emotions.

2.3 Bi-modal Emotion Recognition

Fusion of multiple modalities can achieve better recognition performance than single modality. Nevertheless, a good fusion strategy must be applied; otherwise, the fusion of modalities can hurt the accuracy of the recognition system.

"Gunes and Piccardi studied this case precisely and conducted experiments where only single modalities were tested (facial expression or body gestures) and where both modalities were fused to formulate a detection [24, 28]. The results revealed that the bimodal approach had better performance" [4].

The bi-modal approach is also considered by [7] to recognize emotions by taking into account facial expression and body movements. *"They used neural networks on their solution and achieved much higher average accuracies on fusing both modalities than testing each modality separately, going from* $57.84 \pm 7.7\%$ *on body motion and* $72.70 \pm 3.1\%$ *on facial expression, to* $91.30 \pm 2.7\%$ *average accuracy on bimodal emotion recognition"* [4].

The research proposed by [6] fused the audio-visual, face, and body modalities using the compact bilinear pooling (MCB) method and demonstrated the state-of-the-art results. In this article, besides general feature-level fusion techniques such as average fusion, product fusion, we have also explored the bilinear pooling technique for face and body fusion.

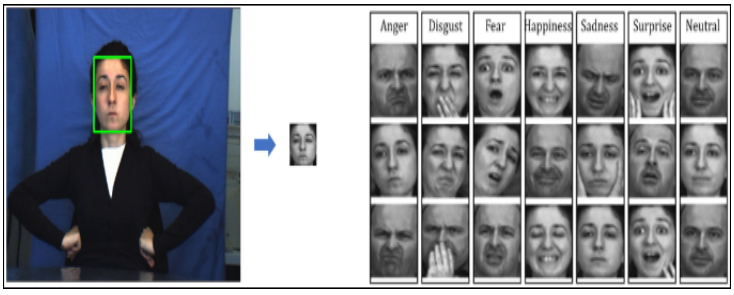


Fig. D.2: Selection of the facial region on a frame from the FABO dataset and scaling into 48*48 to form image database.

3 Proposed System

Our model is comprised of Convolution Neural Networks (CNNs) to extract facial features and bodily features, with linear addition of Long Short Term Memory (LSTM) model to use the sequential information. Each modality (face or body) is trained with CNN in frame-based and sequence-wise to generate the emotional states. It is challenging to determine the multimodal fusion efficiency for emotion recognition accuracy, so we have proposed three different fusion techniques to identify the best approach. The network structure remains the same for all considered fusion modalities. Overview of the system is illustrated in Fig D.1

3.1 Convolutional Neural Network

Convolutional Neural Network (CNN) performs remarkably good at acquiring spatial information. Each CNN layer operates twofold; filtering through the convolution layer and max-pooling to avoid losing useful information. Generally, CNNs are composed of convolutional layers, and fully connected layers extract features. Most of the parameters are also present in the fully connected layers responsible for most of the computation power. For instance, fully connected layers of VGG16 contains 90% of all the parameters. The VGG16 is a deep convolutional network with up to sixteen layers (thirteen convolutional layers and three fully connected layers). Inception V3 reduces the parameters by the introduction of global average pooling [35]. Similarly, Xception [36] takes advantage of the use of residual modules and depth-wise separable convolutions.

To lessen the computation cost, we have implemented CNN architecture as proposed by [37]. *"It is a simple architecture that achieves almost state-of-the-art performance classifying emotions"* [4]. The architecture classifies emotions based on facial expressions and faces according to gender. On the contrary to [37], only relying on the FER-2013 dataset, we use the FABO dataset for training purposes. Besides, our CNN architecture contains four residual deep separable convolutional layers, where batch normalization and ReLU activation function accompany each convolutional layer. *"Batch normalization normalizes the activation of the previous layer at each batch"* [4]. Residual modules modify the desired mapping between two subsequent layers by connecting the output of previous layers to the output of new layers [4]." *Depth-wise separable convolutions reduce further the number of needed parameters. They are composed of depth-wise convolutions and point-wise convolutions "* [4]. Instead of the fully connected layers, this architecture uses Global Average Pooling that reduces each feature map to a scalar by calculating the average of all the elements in the feature map. The last convolution layer has the same number of feature maps as the number of classes. In the end, a softmax activation function is applied to produce a prediction. [4]

3.2 Long Short Term Memory Networks (LSTMs)

LSTMs are a recurrent neural network that processes and absorb sequential information. According to [38]:

"The LSTM states are controlled by three gates associated with forget (f), input (i), and output (o) states. These gates control the flow of information through the model by using point-wise multiplications and sigmoid functions σ, which bound the information flow between zero and one."

To take advantage of spatial and temporal information, we have linearly combined the CNN and LSTM model, where CNN extracts the features and then sequentially feeds into the LSTM network. Such a combination works well in the case of video data, as exhibited by [23, 39].

3.3 Fusion Methods

"One of the issues in multimodal emotion recognition is deciding when to combine the information. There are a few different techniques to fuse the emotion recognition results of different modalities with certain advantages and disadvantages. Some of the most explored techniques are early (feature-level) fusion and late (decision-level) fusion". Recent studies explore another feature-based fusion method called bilinear pooling fusion.

Feature-level fusion *"Feature-level fusion combines the data from both modalities before classification. A single classifier is used containing features from both modalities"* [4]. One of the biggest drawbacks of feature-level fusion is high-dimensional feature production resulting in more parameters and more computation power consumption. To reduce the dimensions, we have applied the compact bilinear pooling (MCB) as proposed by [40]. Bilinear pooling multiplies two vectors that produce tons of parameters, and it is costly. However, compact bilinear pooling reduces the dimensions with the same information level but with very few parameters.

Compact Bilinear pooling fusion [40] proposes a compact bilinear pooling technique for fine-grained visual recognition. In this technique, outer product $\otimes$ is calculated by element wise multiplication of two input feature vectors $f_1 \in V^{n1}$ and $f_2 \in V^{n2}$ and scaling it into a matrix [ ] to reduce dimensions. For instance $y = X\,[f_1 \otimes f_2]$, where X is a learned model, $\otimes$ denotes the outer product and [ ] represents linearizing the matrix in a vector. This technique has produced better results for multimodal emotion recognition task as mentioned by [6], who fused audio-visual, face and body features to recognize emotions by considering co-relation among them.

Decision-level fusion Decision level fusion does not produce high-dimensional features as each modality is trained and classified separately to fuse recognition accuracy at the end. However, this method fails to understand the correlation between input modalities. This co-relation is more important and meaningful in human-machine interactions where body movements and facial expressions complement each other [7].

4 Experimental Results

In this section, we first describe the databases involved and their training protocols. Then we demonstrate the results.

4.1 Benchmark Datasets

We have used the bi-modal face and body FABO dataset and the FER-2013 dataset for the emotion recognition task. Details of the datasets are mentioned as follows:

FABO-Dataset The bi-modal face and body data set is presented by [24]. Two cameras acquire the database for monitoring face and body movements, that captures the facial data and upper body movements separately. The videos provide annotations on the stages of the affective states, therefore splitting the demonstration of each emotion into neutral, onset, apex, and offset phases. Annotation is performed for 16 subjects out of the 23 subjects for emotional classification. The face and body posture tend to shift in the onset process, and these changes reach a steady level at the apex phase. Finally, expressions and movements exhibit relaxation at the offset stage. However, these phase annotations are only done for twelve of the subjects.

Frames in the apex phase are considered for CNN training since they are the ones that reflect the emotions best. Two apex phases are assessed from the annotated videos. The dataset contains 1410 images for anger, 458 for disgust, 343 for fear, 613 for happiness, 570 for sadness, and 588 for a surprise, split into test and training. *"The neutral emotion is the exception, the images for this emotion were obtained from the neutral phase from each video, amounting to 786 images. The selected images display the upper body of the subjects; therefore, a facial recognition algorithm was applied to extract only the facial region within the image"* [4]. The method used was a DNN face detector module included in OpenCV 3.6 . The selected frames were grayscaled and resized to the FER-2013 dataset size, which is 48*48 pixels as demonstrated in the figure D.2

FER-2013 Dataset The FER-2013 database consists of approximately 36,000 images, labeled with seven emotion classes (six Ekman emotional states plus neutral expression). FER-2013 is one of the biggest databases for FER in-the-wild environment but with a low image resolution of 48 * 48 pixels leading to problems for facial landmark detectors. *"The dataset contains 35887 annotated images, with 4953 anger images, 547 disgust, 5121 fear, 8989 happiness, 6077 sadness, 4002 surprise, and 6198 neutral images. Some samples of the images are shown in Fig. D.3"* [4].

Experiments are performed to detect emotions from face and and upper body separately and fused accuracy is also calculated. Details are provided in the section 3.3.

4.2 Network Training

The CNN architecture is trained with benchmark datasets FER-2013 and FABO datasets to extract the facial and body features and evaluate the effectiveness of each modality.

Face-CNN Model: (For Facial Emotional Recognition) Only facial features are trained to the network to evaluate facial expressions. To recognize emotions, first face is localized, tracked, and then face cropping is applied according to the network input parameters. CNN is trained with the FER-2013 dataset, with data augmentation techniques applied to train with more diverse data. It also helps to prevent overfitting and to generalize the model.

Early stopping is used to avoid overfitting. It stops the training process of the model when the error on the validation set gets higher than before. *"The learning rate is reduced when validation loss has stopped improving"* [4].

"The CNN was trained using Adam optimizer. This optimization algorithm is an extension of the stochastic gradient descent. It has some benefits compared to other algorithms, such as less memory requirement, computationally efficient, and it is well suited for problems with extensive data and parameters [41]" [4]. The trained model that we called the face-CNN model achieved 65% accuracy in the validation set. To recognize the emotions from upper body movements, we have trained the CNN model with body features of the FABO dataset.

Body-CNN Model: (For Upper Body Emotion Recognition) Only the FABO dataset is used to train the body-CNN model. This dataset is already described in Section 4.1. However, to train this model, full image frames of the FABO dataset with facial and body gestures information are used, as illustrated in Fig. D.2.

CNN-body architecture is the same as CNN-face, but the images are descaled from their original 1024x768 dimensions to 128x96 and grayscaled to ease and fasten the training process. The pixel values were normalized to a range between -1 and 1. Data augmentation strategies are also applied to increase the number of training samples. Furthermore, the data was split into train and validation data with an 80/20 ratio. The model achieved a 96% accuracy in the validation set.

4.3 Bi-modal Emotion Recognition

As mentioned in section 2, the fusion of different modalities is capable of achieving more significant results than single modalities for emotion recognition. To identify which fusion technique works best in our task, we have applied MCB fusion at the

Fig. D.3: FER-2013 sample images for facial emotion recognition.

Table D.1: Results of different evaluation metrics for each frame-based emotion recognition method.

Evaluation Matrix	Facial Expressions	Upper Body Movement	Bimodal Average Fusion	Bimodal Product Fusion	Bimodal Bilinear Pooling
Precision	77.2 %	72.8 %	81.7 %	82.6 %	83.7 %
Recall	73.0 %	72.7 %	80.4 %	80.9 %	81.5 %
F1-Score	72.8 %	71.5 %	80.3 %	80.9 %	82.5 %
Accuracy	77.7 %	76.8 %	85.7 %	86.6 %	87.2 %

feature-level, whereas product and average fusion strategies are applied at the decision level.

4.4 CNN Architecture

For our experimentation, we have used the same architecture as proposed by [37], as described in detail in section 3.1. However, our implementation varies with [37] as we have used different datasets for training with the addition of the LSTM model to exploit the temporal information. We aim to reduce the number of CNN parameters and computational costs and achieve better generalization. The network is composed of 4 convolutional layers and ReLu and batch normalization at each layer. As mentioned in section 3.1 instead of fully connected layers, global average pooling is applied. However, in the case of a temporal database LSTM model is installed, followed by the softmax.

4.5 Performance Analysis

Frame-based Emotion Recognition

The performance of each modality is tested with our trained network for emotion recognition. Various classification models with different evaluation metrics analyze which modality has better performed. We have analyzed the system performance with precision, recall, accuracy, and F1-score metrics, as illustrated in Table D.1. Moreover, bi-modal fusion with different fusion methods is applied to identify the performance of fusion strategies.

Facial Expression Analysis The normalized confusion matrix in Fig. D.4 shows the percentage of image samples that are of a specific emotion (true label) and that are classified as corresponding to a specific emotion (predicted label). The confusion matrix shows the best classification results corresponding to the anger and neutral emotions with 94% and 90%, respectively. The worst classification result corresponds to the surprise emotion that is often mistaken with fear; however, fear rarely is misclassified as a surprise.

Upper Body Movement Analysis Fig. D.5 displays the normalized confusion matrix for the upper body movements emotion recognition. This confusion matrix shows the true positive rate; hence, the percentage of samples from each dataset that are classified correctly.

From Fig. D.5 it is possible to observe that, once again, the best recognition results are attributed to anger and neutral emotions. *"Comparatively to the facial expression recognition, in this recognition modality, the surprise emotion has a much better recognition rate and is less often mistaken by fear. Also, the sadness emotion is quite often misclassified as a surprise"* [4].

Bi-modal Analysis Two different decision-level fusion methods are tested, an average method and the product-method. *"In the average method, the average is calculated between both modalities and for each of the emotions. In the product method, the product of the probabilities of each modality is calculated for each of the emotions* [4].

Finally, the combination of both modalities produces the results in Fig. D.6 using the average fusion method, and Fig. D.7 using the product fusion method.

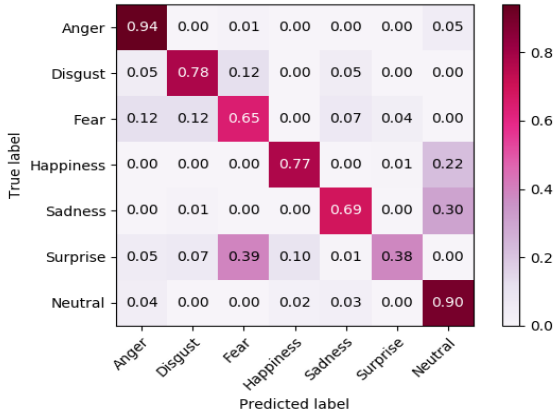


Fig. D.4: Normalized confusion matrix for facial expression recognition.

Sequence-based Emotion Recognition

To train the system with spatial and temporal information of the FABO dataset, we have to use the face-CNN model pre-trained on the FER-2013 dataset. As the FABO dataset posses video data with emotional annotation for 16 subjects out of 23. Each video displays the same emotion from two to four times, so we have divided each video into neutral, onset, apex, and offset maximum phase length of five seconds. We trained this network with these video data and obtained the features, as we have used the full frames of the FABO dataset containing facial and gesture features. These features are feed into LSTM in a timely manner to evaluate the sequential information for emotional classification. Details of recognition accuracy is presented in the Fig. D.8.

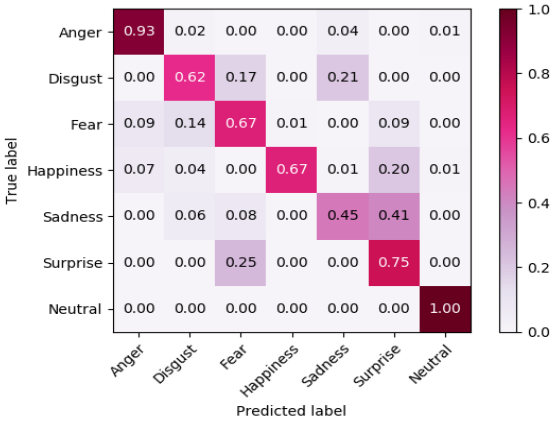


Fig. D.5: Normalized confusion matrix for upper body movements emotion recognition.

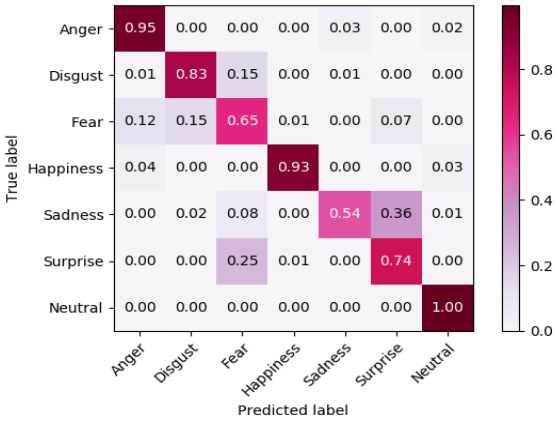


Fig. D.6: Normalized confusion matrix for bimodal emotion recognition using the average fusion method.

Facial Expression Analysis Our network achieved an average accuracy of 93.213 % when it is trained to 80 epochs. It is observed that system performance reached its maximum accuracy when epochs range from 40 to 50; after that system, performance did not fluctuate considerably.

Upper Body Movement Analysis It is observed that temporal information contributed to accuracy efficiency when the network is trained with full FABO dataset frames. Our network achieved an accuracy of up to 79.27 % for emotional recognition through upper body movement analysis.

Bi-modal Analysis When the network is trained with combined facial and upper body features, our system has surpassed the state-of-art accuracy to 94.418 %. In this experiment, network parameters are less than the state-of-art method [6], and it is

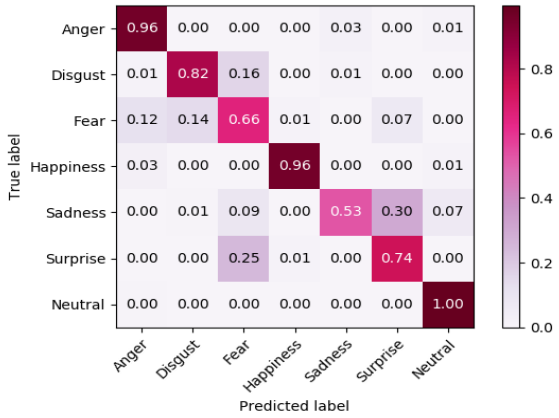


Fig. D.7: Normalized confusion matrix for bimodal emotion recognition using the product fusion method.

robust to work in real-time scenarios.

4.6 Parameters Evaluation

Our network contains four deep-separable convolutional neural layers with ReLu and batch normalization function. We have used global average pooling and softmax for emotion classification that contribute to approximately 600,000 parameters. We trained this network with the FER-2013 database that provides an accuracy of 65% on the validation set. However, usage off the shelf CNN network that is 70 times more parametric heavy and provides 71.3% accuracy on the FER-2013 dataset. In contrast, when we have employed this network to recognize emotions from face and body, it surprised the state-of-art method when we have a bi-modal model with temporal information. Additionally, this model also showed improved accuracy in using a single modality such as the upper body movements. We acquired an accuracy of 76.8 % and 79.27 % with spatial and temporal information, respectively. Application of compact bilinear pooling (MCB) contributed to dimensionality reduction without compromising on the performance.

5 Discussion and Conclusion

The major problem in developing a human-machine affective system is the integration of multimodal sensory information. In this research article, we have explored the spatial-temporal technique for emotion analysis of visual modalities. We have also studied different fusion techniques with lesser computation cost. We have developed a robust architecture to identify emotions from the face and upper body movements to use in real-time human-machine interaction systems.

It is illustrated through the confusion matrices that the bimodal approach shows better results than the monomodal approaches, regardless of the fusion method. In

Table D.2: Performance analysis of our system with CNN and (CNN + LSTM) models and their comparison with state-of-art methods. We have performed 3-fold cross validation after splitting data into 80/20 protocols

Method / Modality	Facial Expressions	Upper Body Movement	Bimodal Fusion
[29]	35.2 %	73.1 %	82.7 %
[25]	66.5 %	66.7 %	75.0 %
[7]	72.7	57.8	91.3
[42]	87.3	74.8	93.65
Our Frame based Model	77.7	76.8	87.2
Our sequence Based Model	90.42	79.27	94.41

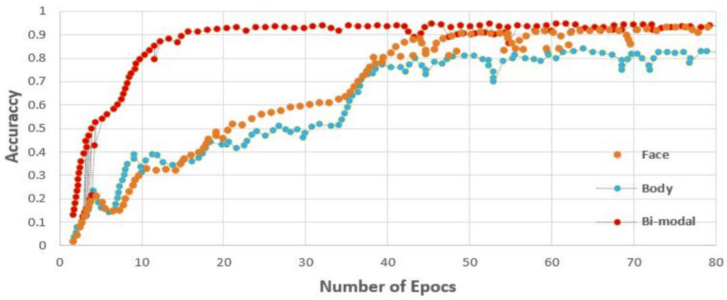


Fig. D.8: Performance of the combined CNN + LSTM neural network model on test data of FABO dataset.

this case, the best recognition rates correspond to both fusion methods for anger, happiness, and neutral emotions. The worst recognition rate is attributed to the sadness emotion that is often misclassified as a surprise emotion.

All the evaluation metrics being considered to have greater values with this approach. Accuracy shows a significant improvement from 77,7% and 76,8% on the facial and upper body movements emotion recognition, respectively, to 85,7% and 86,6% on the fusion of both modalities.

Furthermore, the product fusion method shows slightly better results on all the evaluation metrics than the average fusion method. However, the MCB method surpassed decision level recognition accuracy. It shows that inter modalities relationship towards emotion identification is an essential factor to consider. We have demonstrated that spatial-temporal information is better classified for anger, happy and neutral emotions for further analysis. With upper body movement alone, state of the art methods found it challenging to classify the emotions accurately. However, our system has performed better with the rest of the methods, as illustrated in Table D.2.

6 Conclusions

The major problem in developing a human-machine affective system is the integration of multimodal sensory information. In this research article, we have explored the spatial-temporal technique for emotion analysis of visual modalities. We have also studied different fusion techniques with lesser computation cost. We have developed a robust architecture to identify emotions from the face and upper body movements to use in real-time human-machine interaction systems.

It is illustrated through the confusion matrices that the bimodal approach shows better results than the monomodal approaches, regardless of the fusion method. In this case, the best recognition rates correspond to both fusion methods for anger, happiness, and neutral emotions. The worst recognition rate is attributed to the sadness emotion that is often misclassified as a surprise emotion.

All the evaluation metrics being considered to have greater values with this approach. Accuracy shows a significant improvement from 77,7% and 76,8% on the facial and upper body movements emotion recognition, respectively, to 85,7% and 86,6% on the fusion of both modalities.

Furthermore, the product fusion method shows slightly better results on all the evaluation metrics than the average fusion method. However, the MCB method surpassed decision level recognition accuracy. It shows that inter modalities relationship towards emotion identification is an essential factor to consider. We have demonstrated that spatial-temporal information is better classified for anger, happy and neutral emotions for further analysis. With upper body movement alone, state of the art methods found it challenging to classify the emotions accurately. However, our system has performed better with the rest of the methods, as illustrated in Table D.2

References

[1] A. Augello, F. Dignum, M. Gentile, I. Infantino, U. Maniscalco, G. Pilato, and F. Vella, "A social practice oriented signs detection for human-humanoid interaction," *Biologically inspired cognitive architectures*, vol. 25, pp. 8–16, 2018.

[2] K. Kim, Y.-S. Cha, J.-M. Park, J.-Y. Lee, and B.-J. You, "Providing services using network-based humanoids in a home environment," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1628–1636, 2011.

[3] R. Sorbello, A. Chella, C. Calí, M. Giardina, S. Nishio, and H. Ishiguro, "Telenoid android robot as an embodied perceptual social regulation medium engaging natural human–humanoid interaction," *Robotics and Autonomous Systems*, vol. 62, no. 9, pp. 1329–1341, 2014.

[4] A. R. V. Nunes, "Deep emotion recognition through upper body movements and facial expression," 2019.

[5] C. Breazeal, "Emotion and sociable humanoid robots," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[6] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition," *Computer Vision and Image Understanding*, vol. 174, pp. 33–42, 2018.

[7] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks*, vol. 72, pp. 140–151, 2015.

[8] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[9] L. Y. Mano, B. S. Faiçal, L. H. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, P. Geraldo Filho, G. T. Giancristofaro, G. Pessin, B. Krishnamachari *et al.*, "Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition," *Computer Communications*, vol. 89, pp. 178–190, 2016.

[10] K. Lang, M. M. Dapelo, M. Khondoker, R. Morris, S. Surguladze, J. Treasure, and K. Tchanturia, "Exploring emotion recognition in adults and adolescents with anorexia nervosa using a body motion paradigm," *European Eating Disorders Review*, vol. 23, no. 4, pp. 262–268, 2015.

[11] B. de Gelder, A. De Borst, and R. Watson, "The perception of emotion in body expressions," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 2, pp. 149–158, 2015.

[12] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert)." *Emotion*, vol. 9, no. 5, p. 691, 2009.

[13] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2011.

[14] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE, 2011, pp. 410–415.

[15] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[16] A. Martínez-Rodrigo, R. Zangróniz, J. M. Pastor, J. M. Latorre, and A. Fernández-Caballero, "Emotion detection in ageing adults from physiological sensors," in *Ambient Intelligence-Software and Applications*. Springer, 2015, pp. 253–261.

[17] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.

[18] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[19] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[20] A. Mehrabian *et al.*, *Silent messages*. Wadsworth Belmont, CA, 1971, vol. 8, no. 152.

[21] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2011.

[22] B. Sun, S. Cao, J. He, and L. Yu, "Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy," *Neural Networks*, vol. 105, pp. 36–51, 2018.

[23] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[24] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 1148–1153.

[25] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," *Image and Vision Computing*, vol. 31, no. 2, pp. 175–185, 2013.

[26] C. Shan, S. Gong, and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures." in *BMVC*, 2007, pp. 1–10.

[27] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition," in *Artifical intelligence for human computing*. Springer, 2007, pp. 91–112.

[28] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[29] ——, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 64–84, 2008.

[30] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 619–623.

[31] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *arXiv preprint arXiv:1402.5047*, 2014.

[32] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 106–118, 2011.

[33] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[34] P. Barros, G. I. Parisi, D. Jirak, and S. Wermter, "Real-time gesture recognition using a humanoid robot with a deep neural architecture," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 646–651.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[36] F. Chollet, "Xception: deep learning with depthwise separable convolutions. corr abs/1610.02357 (2016)," *arXiv preprint arXiv:1610.02357*, 2016.

[37] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

[38] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[39] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 451–458.

[40] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnns for fine-grained visual recognition," in *Transactions of Pattern Analysis and Machine Intelligence (PAMI)*, 2017.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adaptive behavior*, vol. 24, no. 5, pp. 373–396, 2016.

Part III

Human Robot Interaction

Paper E

Deep Transfer Learning in Human-Robot Interaction for Cognitive and Physical Rehabilitation Purposes

Chaudhary Muhammad Aqdus Ilyas, Matthias Rehm, Kamal Nasrollahi, Yeganeh Madadi, Thomas B. Moeslund and Vahid Seydi

The paper has been submitted in the
Journal of Pattern Analysis and Application (PAA). Special issue on Computer Vision and Machine Learning for Health Care Applications 2020.

The layout has been revised.

Abstract

This paper presents the extraction of the emotional signals from traumatic brain-injured (TBI) patients through the analysis of facial features and implementation of the effective emotion-recognition model through the Pepper robot to assist in the rehabilitation process. The identification of emotional cues from TBI patients is very challenging due to unique and diverse psychological, physiological, and behavioral challenges such as non-cooperation, facial/body paralysis, upper or lower limb impairments, cognitive, motor, and hearing skills inhibition. It is essential to read subtle changes in the emotional cues of TBI patients for effective communication and the development of affect-based systems. To analyze the variations of the emotional signal in TBI patients, a new database is collected in a natural and unconstrained environment from eleven residents of a neurological center in three different modalities, RGB, Thermal and Depth in three specified scenarios performing physical, cognitive and social communication rehabilitation activities. Due to the lack of labeled data, a deep transfer learning method is applied to efficiently classify emotions. The emotion classification model is tested through closed-field study and installment of a Pepper robot equipped with the trained model. Our deep trained and fine-tuned emotional recognition model composed of CNN-LSTM has improved the performance by 1.47% on MMI, and 4.96% on FER2013 validation data set. In addition, use of temporal information and transfer learning techniques to overcome TBI-data limitations has increased the performance efficacy on challenging dataset of neurologically impaired people.

Findings that emerged from the study illustrate the noticeable effectiveness of SoftBank's Pepper robot equipped with deep trained emotion recognition model in developing rehabilitation strategies by monitoring the TBI patient's emotions. This research article presents the technical solution for real therapeutic robot interaction to rehabilitate patients with standard monitoring, assessment, and feedback in the neuro centers.

1 Introduction

It is challenging for people with traumatic brain injury (TBI) to communicate and socialize due to motor, hearing, and speech inhibitions. For rehabilitation and training purposes, TBI-patients are often treated in specialized neuro centers. Since 2015, our researchers have been working with a national neuro center with a focus on providing technical systems enhancing capability for the residents and to provide assistance and facilitation to staff members [1–3]. The majority of the residents at the neuro center possess unique and highly diverse nature of impaired cognitive and behavioral abilities (for instance, apraxia and aphasia). As some of these residents are unable to recover from their life-altering impairments fully, the center provides full-time care and aid in organizing and supporting activities of daily living (ADL). Providing such facilities is resource, labor, and expertise expensive. It also produces extra strain on the staff members to maintain the same level and standard of services to these residents. One technical means of lifting this burden is intelligent augmented and assistive technologies (AAT) that can be of help to maintain the quality of services and to facilitate staff members in developing and implementing rehabilitation strategies.

Researchers focus on providing assistance in naturals environments through ambient assisted living (AAL). AAL contributes in wide utility space such as from pa-

tients to social services, health workers to smart homes and multi-agent systems with the aim to present a solution for independent living in the user's preferred living environment [4]. AAL aims to provide better quality of life for both elderly people and their care-workers. Recent advances such as the adoption of Internet-of-Things (IoT), cloud computing (CC), virtual and augmented reality (VAR), ambient intelligence (AmI) and neurorobotics have tackled the AAL solutions. According to [5], IoT technologies in the AAL domain are capable of catering to challenges related to ADL, elderly-care, social dis-cohesion, personalized medication, physical activities, health tracking and various other applications. In addition, brain computer interface (BCI) systems contribute to improve the quality of life of elderly people by receiving and transmitting brain signals to external aids and VAR devices [6]. However, the major limitation of employing BCI systems involves wearable sensors mounted on the head to communicate signals to the linked devices, which restricts natural movement of the subjects under observation.

In the AAL domain, researchers have developed specialized AAT systems tailor-made for completion and facilitation of specific tasks such as robots for surgical-operations [7, 8], healthcare robots for monitoring elderly people [9], social assistive robots for social engagement e.g. for children with Autism Spectrum Disorders (ASD) [10–12], or human-computer interfaces for assistance in daily tasks [9]. The AAT systems, specifically developed for elderly care or disabled people, employ different input signals to process information like audio, video, proximity, touch, and their combination is based upon the system application and environment. Over the past few decades, researchers are exerting special efforts to develop such systems with more human-like characteristics like social assistive robots (SAR), to assist in ADL. SARs can be integrated with emotional signal recognition and synthesis for natural and human-like interaction.

There are various ways to extract emotional signals, as one of the regions of the brain stem cell (amygdala) is mainly responsible for generating actions related to emotional arousal [13]. We can identify the activation of signals through this brain region by reactions visible through external and internal body stimuli. For instance, the amygdala regulates the release of hormones in the bloodstream, controls the heart rate, blood pressure, skin conductance, as well as changes in facial expressions [14]. In a nutshell, we can determine these emotional cues by dilation of eye-pupil, electron flow on the skin (skin conductance), brain activity (Electroencephalography (EEG)), Magnetic resonance imaging (MRI), heart rate (Electrocardiography (ECG)), and facial expression recognition (FER) [15] as demonstrated in Fig.E.1. Many researchers focus on the various techniques for the rehabilitation of physical and cognitive impaired people, e.g. [16] establish a virtual reality exposure therapy (VRET) for managing stress reactions. Similarly, [17] develop a BCI system for the extraction of psychological signals of mentally impaired people using electroencephalography (EEG). For developing an affect-based system for use in rehabilitation settings, the real challenge thus lies in the acquisition of emotional signals from people suffering from neurological disorders like patients with acquired brain injury.

Considering the challenges associated with this user group such as limited muscle movement or paralysis, non-cooperative behavior, inappropriate responses, impaired reasoning, involuntary head, and upper body movements, mental inflexibility with

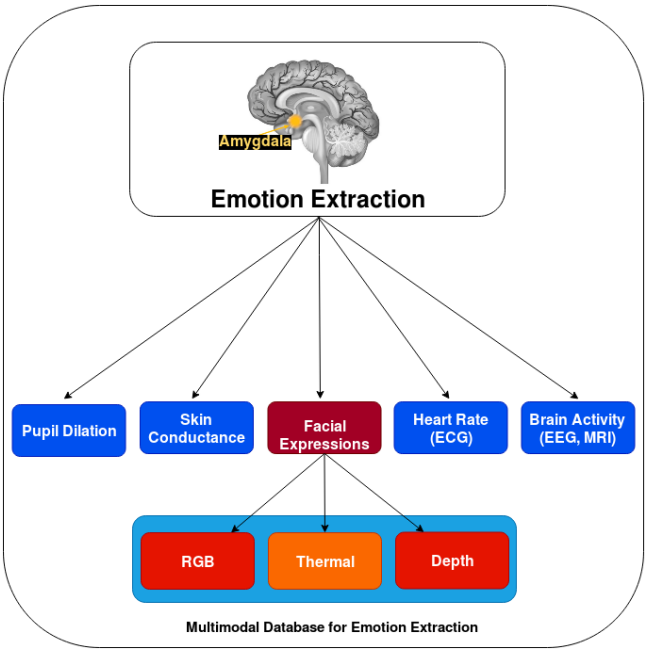


Fig. E.1: Emotional signal identification through various parameters; Collection of data through multi-modal channels to analyze facial expressions

non-compliance, agitation, loud verbalization and sometimes physical aggression, we decided to collect emotional signals in an unobtrusive manner through facial expression analysis [2, 3, 18]. Other methods to identify emotional signals have certain limitations like the installment of sensors on the body e.g., for identifying emotions through skin conductance, ECG, MRI, and EEG. Pupil dilation measurements involve an eye-tracking camera that must be placed close to the face without any occlusion, which is not possible due to limitations related to the physiology of the residents.

Therefore, considering the challenges mentioned above and complexities associated with TBI patients and aiming at capturing data in the natural environment, we extracted emotional signals through facial expressions relying on Ekman's definition of basic emotions. Ekman et al. described six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) as universal basic emotional cues among humans [19]. Details of the data acquisition system in the specified scenarios, modified strategies for improved data quality and pre-processing techniques are mentioned in section 3.

The automatic recognition of facial expressions and interpretation as emotional cues can be utilized in a broad spectrum of socially and emotionally sensitive systems such as robots and virtual humans that engage with people in real-world contexts naturally. Since real-world frameworks encompass uncontrolled settings, where businesses operate in continuous altering circumstances such as occlusions, noise, illumination variations, diverging facial postures, and unwanted head and body movements. Therefore, systems that execute automatic analysis of human emotions must

be robust to visual-data acquisition conditions, varying contexts, and the time of response.

In the past few decades, the performance of automatic-facial expression recognition (A-FER) systems was limited to controlled conditions and posed expressions. These systems were exploiting facial information that is captured in lab environment with majority of induced expressions such as Cohn-Kanade database [20], Cohn-Kanade extended database [21], MMI database [22], JAFFE database [23], DISFA database [24], and DISFA extended database [25]. Therefore, less prone to environmental challenges like illumination and occlusion, pose variation, and spontaneous expressions. Recently researchers are exerting extra effort to develop systems that could perform A-FER in natural circumstances. For this purposes, scientists have collected database in-the-wild such as AFEW [26], SFEW [27], FER2013 [28], ExpW [29], and BU-3DFE and BU-4DFE databases [30, 31]. In addition, emotion recognition challenges are carried out to address the challenges in real-world scenarios. However, researchers have illustrated that facial expression (FE) in naturalistic interaction thoroughly vary from the induced or posed ones [18, 32–34]. Additionally, facial expressions of the TBI patients have additional artifacts such as facial paralysis, non-cooperation during data acquisitions, and large pose variation [3]. Therefore, these databases have the following limitations:

- The databases collected under controlled environmental conditions, with proper illumination and cooperative subject, contain frontal postures in the majority of images or minimal pose variation. However, acquiring data from real-life patients, suffering from brain injuries, is remarkably complex as patients are not cooperative, and it is quite difficult to have frontal postures. Moreover, facial databases captured in-the-wild have diverse features as compared to database captured in the lab environment. Therefore, FER systems trained under "controlled conditions" do not perform well in real-world applications. So it is essential to build a database of TBI patients in natural and unconstrained circumstances.
- Facial expressions of TBI patients significantly vary as compared to healthy people due to prolonged disabilities, paralysis, and continued state of depression. Researches have associated the dominance of negative expressions with this user group [35]. Furthermore, existing FER databases have induced expression, that is different from natural expressions produced involuntary [18, 32–34].
- Some of the TBI patients have additional complexities due to facial paralysis, so their expressions are quite hard to extract. In addition, some of the facial-symmetry and facial bones of the TBI patients are misaligned due to the stroke. Images with such features are not available in current databases.
- Facial expressions of healthy people are easily distinguishable such as happiness, sadness, anger, fear, surprise, disgust, and neutral. TBI patients do not have clear six expressions, but we find a prominence of only two to three expressions, usually the negative ones. It is essential for deploying affect-based intelligent interactive systems with these users that systems are trained on a specially dedicated database, developed in real environmental conditions with all the complexities associated due to the brain injury and real-world challenges.

In this paper, we aim to address the limitations mentioned above by the development of a TBI patient database under natural, unconstrained, and uncontrolled conditions. This multimodal visual database is collected with RGB, thermal, and depth sensors in the specific scenarios to ensure uniformity and reliability in data collection. Database annotation is performed by the neuro center staff members, experts, caregivers, physiotherapists, and doctors, who worked with a particular resident for more than six months. It contains a range of expressions from the residents performing daily activities like physiotherapy, cognitive rehabilitation activities, and social communication. We have collected 1723 videos in 91 sessions, illustrating emotional reactions of 11 subjects in three modalities: RGB, Thermal, and Depth.

Table E.1: Subjects in database along with challenges due to TBI, number of sessions and activities

Subjects	No. of Sessions	Activities			Challenges			Prominent Features
		Cognitive	Physio	Social	Body Paralysis	Speech Inhibition	Facial Paralysis	
A	12	4	4	4	Complete	Yes	Partial	High Anger
B	10	4	3	3	Left Side	No	No	High Arousal
C	10	4	3	3	Lower Body	No	No	Excessive Head Movement
D	9	3	3	3	Partial	No	Partial	Emotionally Unstable
E	9	2	4	3	No	Yes	Partial	Emotionally Unstable
F	7	2	3	2	Partial	No	No	High Arousal
G	6	2	2	2	Lower Body	No	No	Excessive Upper Body Movement
H	7	2	3	2	No	No	Partial	Low Arousal
I	6	2	2	2	Yes	Yes	Partial	Low Arousal
J	8	2	3	3	No	No	No	Verbal and Physical Aggression
K	7	3	3	1	Partial	Yes	No	Emotionally Unstable

There exists a vast range of emotional and facial expression recognition databases. However, they have limitations, mostly because data is acquired in controlled lab environments. Additionally, all of the existing databases are of healthy people with quite clear expressions that are remarkably different from brain-injured residents of the neuro center, who do not show the same variation in the six basic expressions. To reach more realistic and exact results, we developed the TBI patient database. As we know, learning deep NNs needs massive labeled training data. So we applied a deep transfer learning model to utilize related data from other databases to help the training the model.

The main contributions of the paper are as follows:

- This research article focuses on the extraction of psychological signals of neurologically impaired people using transfer learning (TL) techniques that assist the care-workers to monitor and assess the rehabilitation process with increased emotional efficacy.
- The research article contributes to designing a specialized framework for collecting consistent and reliable data from neurologically impaired people for social, physical, and cognitive well-being.
- We employed a deep architecture of CNN and CNN plus RNN to develop a FER model. This FER model is tested on CK+, MMI, JAFFE, FER-2013, AFEW,

SFEW2.0, DISFA, and ExpW databases and competes with the state-of-the-art methods and outperforms some of them.

- It is demonstrated that the deep trained FER model is capable of recognizing emotions of people with facial paralysis in a natural environment, producing state-of-the-art performances.
- Integrating the FER model with the Softbank Pepper robot to recognize emotions helps the staff members and care workers to understand the emotional conditions of the residents better and adopt the rehabilitation and interaction strategies in real-time.
- Our findings indicate that the robot intervention with the residents of the neurocenter enhanced the productivity of physiotherapy and social interaction.

The rest of the paper is organized as follows. Section 2 provides an overview of existing databases and related research in the field of Facial Expression Recognition (FER) with the focus on natural data collection environment. Section 3 explains the process of data collection of brain-injured patients in various scenarios. Section 4 presents the methodologies implemented in our approach. Section 5 describes the experimental studies and result evaluation. Section 6 illustrates the contribution towards rehabilitation strategies. Section 7 concludes the paper.

2 Related work

2.1 Current Databases

Existing databases of facial expression recognition such as Cohn-Kanade (CK, CK+) [20, 21], MMI [22], CE [36], JAFFE [23], and BU-4DFE [30, 31] are developed in lab and controlled conditions where subjects displayed distinctive facial expressions. These databases have high-quality based posed-expressions. However, non-posed and spontaneous expressions acquired in uncontrolled or in-the-wild environments are quite different from posed expressions. It is essential to identify non-posed expressions in a natural or uncontrolled environment for automatic affective computing. Thus, researchers focused towards data acquisition in-the-wild or uncontrolled settings such as AFEW and SFEW datasets [27], used in series of EmotiW challenges[1], or FER-2013 [28], DISFA [24], DISFA+ [25]. These databases encompass multimodal effects such as voice, biological parameters, and sequences of frames. However, due to the number of subjects, pose variation, and environmental settings, the range of diversification of these databases is minimal. We briefly describe the databases that are captured in-the-wild as well as in controlled settings, used for emotion recognition, and will discuss their limits leading to the creation of the TBI database.

CK+ Database The Extended Cohn-Kanade (CK+) database [21], is one of the most extensively used databases for FER systems. It is established in the lab or controlled settings, with 593 image sequences of 123 subjects, of which only 327 are annotated with seven emotion labels (six basic emotions and contempt). The database

[1] https://sites.google.com/site/emotiwchallenge/

Table E.2: An overview of the facial expression databases

Databases	No. of Sub.	Samples	Env.	Nature (Posed / Spontaneous)	Expressions Information	Availability
CK+ [21]	123	593 Image Sequences	Controlled (Lab)	Posed & Spontaneous	6 Basic expressions (with contempt) plus Neutral	http://www.consortium.ri.cmu.edu/ckagree/
JAFFE [23]	10	213 Images	Controlled (Lab)	Posed	6 Basic expressions plus Neutral	https://zenodo.org/jaffe
MMI [22]	25	740 Images 2,900 videos	Controlled (Lab)	Posed	6 Basic expressions plus Neutral AU-FACS	https://mmifacedb.eu/
DISFA [24]	27	89,000 Images	Controlled (Lab)	Spontaneous	(6 Basic expressions plus Neutral (by EMFACS system))	http://mohammadmahoor.com/disfa
FER2013 [28]	N/A	35,887 Images	Web (In-the-wild)	Posed & Spontaneous	6 Basic expressions plus Neutral	https://www.kaggle.com/fer2013
AFEW [26]	330	1,809 Videos	Movies (In-the-wild)	Posed & Spontaneous	6 Basic expressions plus Neutral	https://sites.google.com/view/emotiw2018/home
SFEW2.0 [37]	N/A	1,766 Images	Movies (In-the-wild)	Posed & Spontaneous	6 Basic expressions plus Neutral	https://cs.anu.edu.au/few/AFEW.html
ExpW [29]	N/A	91,793 Images	Web (In-the-wild)	Posed & Spontaneous	6 Basic expressions plus Neutral	http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html

consists of 69% females and 31% males with an age range from 18 to 50 years. The dataset contains posed and non-posed facial expressions at a maximum intensity level.

MMI Database The MMI database [22] is captured in the laboratory or controlled settings with 326 image sequences of 32 subjects. Two hundred thirteen image sequences are labeled with six basic expressions with onset-apex-offset states.

JAFFE The Japanese Female Facial Expressions (JAFFE) [23] database is captured in controlled conditions. It consists of 213 image samples of 10 female subjects. Each subject has 3-4 facial images with each of six basic expressions and one image with a neutral expression.

DISFA Denver Intensity of Spontaneous Facial Actions (DISFA) database [24] consists of 27 subjects captured with spontaneous expressions. It is coded with Action Units (AUs) ranges from 0 to 5 with zero corresponds to the absence of any activation of muscles, while five belongs to maximum intensities. We have employed the EMFACS conversion system [38] to convert AU FACS codes to emotional expressions that presented approximately 89000 images with a majority having neutral expressions.

EmotiW-AFEW-2018 Acted Facial Expressions in the Wild (AFEW) [27] and its subset Static Facial Expressions in the Wild (SFEW) [37] have been used as a benchmark dataset for annual emotion recognition in the wild challenge (EmotiW) challenge. AFEW is a multimodal-temporal database containing facial expressions from movies and reality TV shows that are close to real-world scenarios. AFEW consists of 330 subjects with an age range from one to seventy-seven years (1-77 yrs). The annotation of this database is according to six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) and a Neutral expression. The AFEW 7.0 dataset used in EmotiW 2017 consists of subject independent data partitions with training (773 samples), validation (383 samples) and test sets (653 samples).

SFEW Static Facial Expressions in the Wild (SFEW) [37] is developed by extracting few images from AFEW with varied head poses, close to real-life illumination

Table E.3: Number of data images for each expression for the databases

Database	CK+	JAFFE	MMI	DISFA	AFEW2018		FER2013		SFEW		ExpW	
Image Size	640*490 720*480	256*256	720*576	768*1024	N/A		48 * 48		720*576		N/A	
F-Exps					Training	Val	Training	Val	Training	Val	Training	Val
Anger	90	30	1959	436	118	59	4953	958	178	77	1272	318
Contempt	36	0	0	0	0	0	0	0	0	0	0	0
Disgust	0	29	1517	5326	72	39	547	111	66	23	1250	312
Fear	50	32	1313	4073	76	44	5121	1024	98	47	329	82
Happy	138	31	2785	28404	142	63	8989	1774	198	73	10576	2644
Neutral	324	30	3034	48582	129	61	6198	1233	150	86	8309	2077
Sad	56	31	2169	1024	104	59	6077	1247	172	73	2494	623
Surprise	166	30	1746	1365	70	46	4002	831	96	57	2471	617
Total	860	213	14523	89210	711	371	35887	7178	958	436	26701	6673

conditions, age-range, and distinctive facial expressions. The SEFW 2.0 is used in the EmotiW 2015 challenge and its most commonly used in general. The dataset is divided into three partitions: Training set (958 image samples), validation set (436 image samples), and test set (372 image samples). Each image sample is assigned with one of seven basic expressions, i.e., anger, disgust, fear, happy, neutral, sadness, and surprise.

FER-2013 The FER-2013 database [28] consists of approximately 36,000 images, labeled with seven emotion classes (six Ekman emotional states plus neutral expression). The database is established by using Google image search combined with phrases for gender, age, ethnicity, and 184 emotion-related keywords. FER-2013 is one of the biggest databases for FER in-the-wild environment but with a low image resolution of 48 * 48 pixels leading to problems for facial landmark detectors.

EXPW The Expression in-the-wild (ExpW) database [29] is comprised of approximately 90,000 facial images downloaded from the web. Each of the images is manually assigned to one of the seven primary expressions.

Nonetheless, all the databases, as mentioned earlier, consist of images of healthy people without any facial paralysis, cognitive or physiological impairments. Hence, there is a need for the development of systems dedicated to cognitive and physical impaired persons like TBI patients, based on natural, spontaneous, unposed, and uninduced facial expressions. To address these demands, we developed a database of TBI residents in natural and uncontrolled settings, details provided in section 3.

2.2 Current Architectures for Affect Recognition

Automatic affective computing is a well-established research area, and there are a wide variety of algorithms and databases to develop automated affect recognition mechanisms. We would like to briefly discuss state-of-the-art methods for emotion-related search on the databases explained in section 2.1. Emotion recognition systems can be distinguished by the methods employed for feature extraction and feature classification. Most of the advanced FER systems are exploiting the techniques of Convolutional Neural Networks (CNN) for facial feature extraction and classification, as they provide state-of-art results for facial expression recognition [39–41], pain iden-

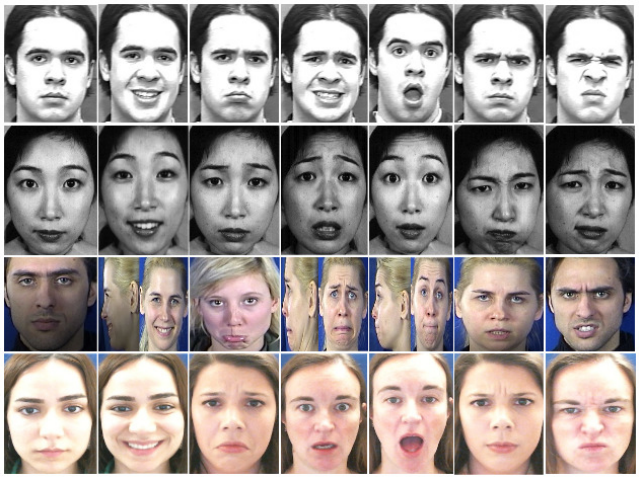


Fig. E.2: Sample database images captured in controlled conditions for facial expressions: Databases (Rows Top to Bottom) CK+ , JAFFE, MMI and DISFA+ ; Emotion categories; (from left to right) Neutral, Happy, Sad, Fear, Surprise, Angry & Disgust (For CK+ contempt)

tification [41] and interpretation as emotional states [42, 43]. Conventional algorithms for affect recognition use handcrafted features such as pixel intensities [44], Gabor filters [45], Local Binary Patterns (LBP) [46, 47], Local Quantized Pattern (LQP) [48] and Histogram of Orientated Gradients (HoG) [49]. Handcrafted features are accompanied by unintended features that have no or less impact on classification. In the case of handcrafted features, not all possible cases can be included for features selection and classification, so its performance is compromised.

The significant advantage of deep learning methods over conventional machine learning models is the simultaneous performance of feature extraction and classification. Moreover, deep learning methods apply iterative approaches for feature extraction and optimize error by back propagation, thus resulting in those critical features that human experts can miss while handcrafting the features. Recently used deep learning algorithms for FE and emotional analysis has demonstrated a remarkable ability to learn features and achieved state-of-the-art results in a range of learning tasks like cross-database evaluation where handcrafted features exhibit low performances due to lack of generalization to new scenarios. Moreover, deep neural networks perform remarkably well for subject independent estimation schemes of emotional expression recognition. This interdependence contributes to the formulation of this paper, as the stability and reliability of the deep learning systems could perfectly align with the procedures required for clarifying complexity in emotion analysis in natural and unconstrained environments, mainly dealing with brain-injured patients.

Deep neural networks, notably CNNs, are well-established approaches for researchers in the field of deep-vision for FER. In the FER-2013 challenge, [70] achieved the 1st prize by exploiting deep neural networks in two stages: use of CNN trained in a supervised way at a first stage and a second stage applying Support Vector Machines (SVM) on the output of the trained CNN. Kahou et al. [71] winner of the EmotiW-2013

Table E.4: Summary of Architectures and Methods for Affect Recognition.

Method	Database	Architecture
Mohammadi et al. [44]	CK+, MMI	Sparse representation classification, PCA based dictionary building
Shan et al. [47]	MMI, JAFFE	Boosted Local Binary Pattern (B-LBP) + SVM
Zhao and Zhang [46]	CK, JAFFE	Kernel Discriminant isometric mapping (KDIsomap)
Liu et al. [50]	EmotiW-2014 (AFEW)	Multiple Riemannian kernels + SVM
Liu et al. [39]	CK+, JAFFE	Boosted Deep Belief Networks (BDBN)
Yao et al. [51]	EmotiW-2015 (Audio-Video) SFEW, AFEW	Emotional expression relation and facial muscle Activation Unit (AU) with RBF kernel
Kaya et al. [52]	EmotiW-2015 (Audio-Video) AFEW, AFEW	Partial Least Squares Regression (PLS) and Kernel Extreme Learning Machines (ELM) with multi-level weighted fusion
Ng et al. [53]	EmotiW-2015	Transfer learning for deep CNN, Pre-trained on the ImageNet dataset; cascading fine-tuning
Yao et al. [54]	Emotiw-2016	HoloNet, CNN with Concatenated Rectified Linear Unit (CReLU)
Rodriguez et al. [55]	CK+	VGG-16 + LSTM
Yan et al. [56]	AFEW6.0, CHEAVD (Audio-video)	Multi-cue fusion; Cascaded CNN and Bi-directional-RNN CNN + SVM
Liu et al. [57]	CK+, MMI, SFEW	CNN with Loss layers
Li et al. [58]	CK+, SFEW,	Deep Locality-Preserving CNN (DLP-CNN)
Zhang et al. [29]	CK+, SFEW, FER-2013	CNN with Multi-task network (MN)
Kim et al. [59]	FER-2013	Discriminative deep CNN (DCNNs); alignment-mapping networks (AMNs); CNN with Network Ensemble
Meng et al. [60]	CK+, MMI, SFEW	CNN with MN; Identity-aware CNN (IACNN)
Yu and Zhang [61]	SFEW	CNN with Network Ensemble
Zhao et al. [62]	CK+	Expression intensity-invariant Network (EIN)
Yu et al. [63]	CK+	Expression intensity-invariant Network (EIN) + Multi-Task-CNN (MN)
Kim et al. [64]	CK+, MMI	Expression intensity-invariant Network (EIN) with data augmentation, illumination normalization and face frontalization
Zhang et al. [65]	CK+, MMI,	Network ensemble with cascaded CNN and SDM
Kuo et al. [66]	CK+	Applied FA network and Intraface
Sun et al. [67]	MMI	NE with GoogLeNet and SDM
Otberdout et al. [68]	AFEW	Deep CNN + Symmetric Positive Definite (SPD) matrices
Fan et al. [69]	AFEW	CNN with VGG-LSTM and fusion techs

Fig. E.3: Sample database images captured in uncontrolled conditions for facial expressions: Databases (Rows Top to Bottom) FER2013, AFEW , SFEW, ExpW ; Emotion categories; (from left to right) Neutral, Happy, Sad, Fear, Surprise, Angry & Disgust

challenge, used the CNN and deep belief network (DBN) composed of two-stacked layers of Restricted Boltzmann Machines (RBMs). The first layer of RBM comprised Gaussian RBM with noisy ReLU, and the second layer Gaussian-Bernoulli RBM. This method worked well and managed to get the at-the-time state-of-the-art performance but at higher computation cost for larger datasets. In 2014, [39] incorporated three tasks of feature learning, feature selection, and classification in a unified manner by employing Boosted Deep Belief Networks (BDBN) and managed to achieve remarkable results in challenging conditions. The winner of the EmotiW-2014 challenge [50] combined multiple kernels on Riemannian manifolds for emotion classification by the measurement of corresponding similarities and distances. Researchers in [50] employed SVM, logistic regression, and least-squares models for emotion classification and applied decision level fusion. However, along with high computation cost for feature extraction, this method produced lower accuracy when exposed to challenging emotional categories.

Kulkarni et al. [72] demonstrated the good results to determine whether 6-class expressions are genuine or these facial movements are fake. He addresses the problem by projecting facial features in deeply learnt space. However, 12 class and the binary emotion pair classification problem still remains a challenge. This is because the distinguishing factors between the unfelt and genuine expressions occur in a very short part of the whole emotion and are a challenge to model. Guo et al. [73] presented dataset with 50 classes of compound emotions for affective computing and geometrically represented the landmark displacement to recognize emotions. However, it is challenging to determine dominant or complementary emotions. Yao et al. [51] explored the significance of the suppressed relationship between evolving characteristics derived from facial muscle motions. The particular relations and patterns between emotional expression and facial muscle Activation Unit (AU) are extracted and called it AU-Aware facial features. This method leads them to surpass the EmotiW-2015 challenge without using additional data. [52] applied two least-squares regressions, specifically Partial Least Square (PLS) and Kernel Extreme Learning Machines (ELM)

with multi-level weighted fusion for emotional classification. one of the drawbacks of applying multi-level fusion with different input modalities audio or video could result in performance downgrading. [53] applied transfer learning techniques on a small dataset for static facial expression recognition in the wild, by pre-training their network on ImageNet dataset followed by fine-tuning to target dataset and achieved comparable results.

In the year 2016, [54] applied a deep but computational efficient CNN with Concatenated Rectified Linear Unit (CReLU) and inception- residual structural for emotional recognition under unconstrained environment. In the year 2017, [55] exercised CNN to learn features from VGG-Faces and integrated with Long Short Term Memory (LSTM) to gain the temporal information. This approach was further improved by [41], who applied deep CNN for features classification into expressions and fed the system with super-resolved facial images. [56] employed the cascaded CNN and RNN, where images are first fed into CNN for facial features extraction followed by bidirectional RNN to learn the changes. One of the common aspects in the work of the [41, 53–55] the use of extensive annotated data of healthy people, captured in controlled and uncontrolled environmental conditions. Transfer learning can be applied to overcome the challenges of training CNNs that require large annotated training datasets of diverse expressions. Transfer learning overcomes the limited data problem by transferring image features learned with CNNs on large datasets to other visual recognition tasks on targeted, limited training data samples [74]. In the case of TBI database, transfer learning is applied to learn features from large-scale public datasets captured in varied environmental conditions and distinct scenarios, with the presence of all expression states, to serve as a better weight initialization by fine-tuning.

As the work in [53, 55, 56, 74], exhibited state-of-the-art results for emotional challenges, but healthy subjects. Therefore, we investigated a similar approach for the TBI dataset. We employed CNN pre-trained to VGG-16 to learn the features from eight public databases and then by applying transfer learning approaches, fined-tuned to TBI dataset to overcome the identity and unbalanced emotional-data limitations.

3 Traumatic Brain Injured People Database (TBI-Database)

3.1 Data Acquisition

Data collected at a neuro center that offers 24/7 rehabilitative care for their residents with brain injury. The goal was to record visual data from the residents in natural scenarios to extract emotional information. Due to the nature of their impairments, it is very complex to collect data for all expressions of anger, sadness, happiness, surprise, and disgust. Moreover, residents have diverse cognitive, physical, and interactive skills. Sometimes the residents demonstrate physical and verbal aggression along with inappropriate responses. Most of the computer vision techniques for FER are dependent on data quality and environmental conditions like occlusion, lighting, and face pose and alignment. Considering these conditions, we collected the data in three different scenarios with the help of experts, trainers, and caregivers to have

reliable and the best possible quality of the data in unconstrained scenarios. These situations are a) cognitive rehabilitation strategies, b) physical rehabilitation strategies, and c) social interaction aiding strategies. Generally, a caregiver follows a set of protocols [75] for the rehabilitation tasks.

In order to deploy automated affect-based systems based on facial expressions, it is necessary to set up a signal perceiving sensors-system, in our design RGB, thermal, and depth sensors. However, there is no extensive research explaining data collection methods for the FE of people that have suffered from TBI residents.

The studies in [22, 76] explained database creation and organization of healthy and cooperative subjects with spontaneous and induced expressions in a controlled lab environment or in-the-wild settings or through online websites. However, in the case of our residents, there is no database, or database development protocols, so we relied on data acquisition with rehabilitation protocols and then modified them after analyzing them carefully. We set up the data acquisition system with RGB, thermal, and depth cameras, placed at 1.5 meters distance from the residents while performing their rehabilitation and social activities. Experts prescribe playing games as a therapy is the most effective way to aid brain injury recovery [77–79]. Researchers recommend five games for brain injury recovery: Card games, Sudoku, Lumosity, TherAppy, and Tetris [80]. We modified these games, including other rehabilitation activities to obtain optimal data for the training of a deep learning-based system; details are provided in the later sections 3.1-3.1

Data collection approaches are distinguished by the rehabilitation activities and the disability of the resident. We collected data from eleven residents. The precise nature of their disability is described in table E.1. Due to severe and diverse conditions of these residents with emotional instability, experts plan strategies for their recovery based on their health conditions and neuropsychological test results [75, 81]. Furthermore, these residents have impaired facial and emotional expressions, accompanied by frequent mood swings, low concentration, and significant pose variations in regards to the capture of facial images.

It is also challenging to extract all six basic expressions, so to have useful facial video data, we altered the standard rehabilitation activities to gather more diverse information.

Cognitive Rehabilitation Strategy

The basic aim of this activity is to improve the ability of residents to understand and interpret information to perform specific functions mentally. Emotional stability is a key factor in this training; otherwise, residents will not be able to participate and get the advantage of these exercises. For this purpose, caregivers follow a set of protocols like Mini-Mental State Exam (MMSE)[2] and Montreal Cognitive Assessment (MoCA)[3] comprised of repetitive activities with gradual increase in difficulty level, to assess the attention, memory [82], visuospatial perception [83], language and communication, function execution and learning ability of brain-injured residents [81]. These tasks

[2]https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/undersoegelser-og-proever/skemaer/geriatri/mms-mini-mental-status/

[3]https://www.mocatest.org/

are mostly accomplished through the use of calendars, drawing clocks, memory log or memory devices, alarms or reminders, reading or listening to books, and playing games. The majority of these activities were performed on the paper placed on a table. During these activities, we encountered a couple of problems that resulted in poor data quality: a) subjects mostly looked downwards, b) frequent pose changes, and c) less attention. Hence, these rehabilitation tasks were tailored to the requirements of the residents in the following ways:

- Residents performed the tasks on a PC tablet, as mentioned earlier, that was placed in parallel to the cameras, which resulted in more frontal facial images and increased attention.
- A favorite movie clip or cartoon character of a resident was displayed on the screen, and then residents were asked about the character or the story. This activity was repeated, and the cognitive assessment was monitored accordingly.
- Error-less (EL) learning was performed by instructing residents to sing lyrics of songs, match pictures, stack Lego bricks, and play computer games, which are of the subjects' interest.
- Sudoku is an organizational game with numbers, colors or alphabets, normally played on paper. Residents played this game electronically on the tablets placed at a predefined location and orientation, resulting in frontal facial images. Most of the residents found the game apparatus comfortable, and there was a wide range of games from easy to hard difficulty providing the opportunity for trainers to monitor the learning skills of the resident at each level.
- Older residents preferred card games rather than playing digitally. Therefore, card games like Memory, Solitaire, Go-fish, and war were played with them. These games proved to be beneficial in recovery as they involve strategy and thought processes with smaller challenges [80]. Regularly playing these games boosted memory skills as well as mathematical understanding, depending on the game. Cognitive skills assessors confirmed this result.
- We have introduced another application based game 'Lumosity' for improved memory, problem-solving, and to speed-up processing. This app presents the range of brain training games based on the input information to improve learning skills. Residents showed a positive response to this app.
- Residents suffering from speech problems were asked to play TherAppy, an application based game developed by Tactus Therapy Solutions, created for residents' language skills recovery. This game comprises of four modules for Comprehension, Naming, Reading, and Writing [84]. Residents were asked to recall the name of a picture, complete a phrase, or spell a word after listening to a short sound clip. Hints were available by clicking a button if a resident was struggling.
- Most of the residents exhibit negative expressions like sadness, depression, anger, or aggression more frequently. In order to have other expressions like surprise, happiness, or joy, various games were created in such a way that intentionally lead to winning for the residents that resulted in more positive expressions.

Attention and memory enhancement are core elements in mental training. All

these modified strategies were implemented on eleven residents, generated less erroneous database, and the residents exhibited more expressions and learning as compared to the custom exercises for cognitive skills recovery. Cognitive skills were evaluated by meeting goals and levels of mental-games applications. Performance evaluation is discussed in detail in section 5

Physical Rehabilitation Strategy

TBI causes physical morbidity due to damage to the sensory-motor system. Depending on the nature of the damage, it can cause reduced muscle movement and paralysis to the upper limb, lower limb, or complete body. Physical rehabilitation methods are planned case to case while considering age, gender, disability type, and post-concussion symptoms [85]. Additionally, assessment of activity tolerance, balance, coordination, and postural control estimation are taken into account while conducting cardiovascular, muscular-skeletal, and vestibular activities. Physiotherapists conduct these activities through preset operations like cardio exercises, using a treadmill, walking or mild running independently or with a trainer, cycling, push-ups, squats, and other related exercises after assessing the abilities of residents [85]. During all these activities, facial data is hardly available due to the excessive movement of the body or face. Therefore, to acquire the maximum facial data, we asked residents who do not have or have partial paralysis to perform physical exercises:

- Residents ride a stationary bicycle to have a static upper body as much as possible while looking at a tablet placed parallel to cameras. During the exercise, expressions were recorded.
- For residents who use wheelchairs, the tasks were designed accordingly, so they moved their chair forward and backward within three meters for multiple sessions.
- Activities such as hand press-ups, arm raises, and cup pick-up and placing were performed.
- Console video games were also introduced, which aided the movement of the resident arms and hands to a certain extent while playing. These games exhibited more explicit expressions and hand-eye coordination.
- Card games also helped with training dexterity and gross motor skills.

These activities resulted in useful data while enhancing the interest of residents throughout the therapy sessions.

Social Rehabilitation Strategy

Social rehabilitation is quite a complex and long-term challenge due to cognitive and behavioral disorders. Social reintegration strategies are based on individual cognitive progress, mental health, and behavioral distortions. In a standard scenario at the neuro center, the residents sit around a table over a cup of tea and share their daily activities. Often, residents do not take an interest, and trainers have to intervene by asking questions. Another observed problem is that residents with speech inhibition communicate through writing letters on tablets, which slows down communication

and reduces interest. To overcome these challenges, we introduced the following activities:

- Firstly, we shared storybooks with the residents and asked them to read aloud to other residents of the neuro-center. Most of the participants did not take an interest in listening to the story due to poor storytelling skills and limited concentration.
- Secondly, we played card games with residents resulting in better interaction with the other participants as compared to the storytelling activity.
- Thirdly, we utilized PS4 console games. Every participant showed interest individually or as part of a team. Most of the participants enjoyed Medal of Honor Airborne (MOHA)[4], Need for Speed[5] and similar games. When playing MOHA in two teams, participants of each team worked closely with each other, enhancing mutual interaction. They also expressed their emotions better at the different stages of the games.

These activities also helped in physiotherapy. However, it is still challenging to get all the emotional states due to non-cooperation, traumatic disabilities, and other social and technical issues; therefore, we have further classified the expressions into positive and negative expressions [3].

Data collected in multiple phases throughout 91 sessions, as presented in table E.1 with RGB, thermal, and depth sensors. In total, we collected 1723 video events, each of a maximum of 5 seconds in length.

3.2 Data Annotation

Furthermore, for accurate annotations, only those experts or trainers were consulted who worked with these residents for more than three months and have at least ten months of experience dealing with residents that suffered from brain injury. Experts annotated the videos manually and then later verified when image sequences are split into various categories. Various pre-processing steps are applied to develop a high-quality facial database; details are provided in section 4.

4 Methodology

In this section, we describe the three main steps for the automatic recognition of facial expressions (FE), i.e., pre-processing, facial feature learning, and facial features classification. The algorithms explored and state-of-the-art implementations for processes, as mentioned earlier, are presented below:

4.1 Pre-processing

Pre-processing is a vital step to avoid unwanted features for facial expression recognition, such as illumination variations, background clutter, and different head poses.

[4]https://www.ea.com/games/medal-of-honor
[5]https://www.ea.com/games/need-for-speed

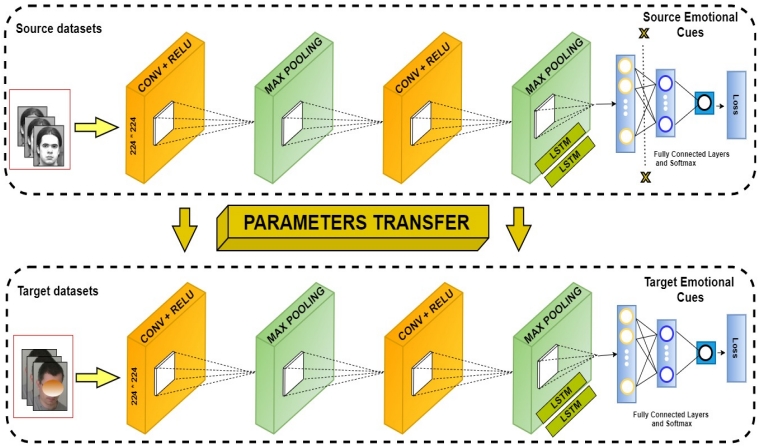


Fig. E.4: Transfer Learning Model Architecture

Therefore, to ensure the learning of only essential features, we applied the following pre-processing algorithms before exposure to neural networks training for the formation of a high-quality facial data log.

Face-alignment

The first step for FER tasks is face detection to remove background and non-relevant features. Viola-Jones (VJ) [86] is a classical method, widely used for face detection that is robust and accurate for frontal faces. However, the algorithm exhibits lower performance in natural and in-the-wild environments, where faces are not always frontal, producing false detection. To achieve higher quality data, we have used the dlib-CNN-face detector [87], that has surpassed VJ for face detection, under unconstrained and natural environmental conditions with significant pose variations [88]. In addition, for further face alignment, we have estimated the facial landmarks through a cascaded regression method, i.e., Supervised Descent Method (SDM), which tracks 49 facial points and reduces the variations and in-plane rotation.

Illumination and Pose Normalization

Deep neural networks are sensitive to illumination and contrast, which can lead to significant intra-class variations even when the images of the same person displaying the same expressions have different contrast and illumination. We have employed histogram equalization combined with illumination normalization, as this method has produced state-of-the-art results in the literature of FER [89]. Another challenge, associated with unconstrained and natural settings, are facial images with large pose variations. We have employed the pose normalization technique that produces frontal views, where landmarks are calculated with arbitrary facial positions, and by finding the inverse of the transpose matrix, the face is frontalized [90].

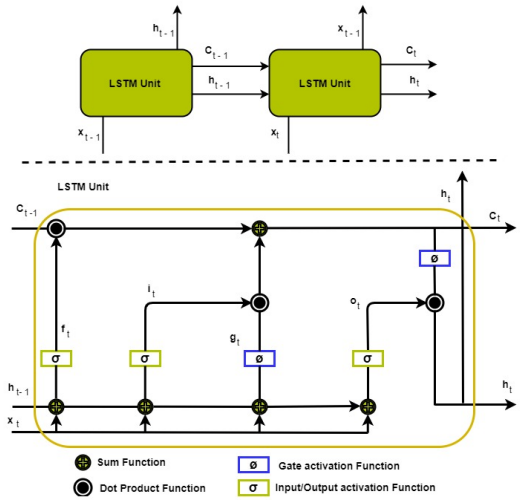


Fig. E.5: LSTM architecture with memory unit

4.2 Deep Learning Architecture for Feature Learning and Transfer Learning (Convolutional Neural Network)

Our work is focused on the emotion cues from images and sequences of images. Convolutional layers are richly embedded with spatial information. We have used the features from convolution layers instead of fully connected layers and transferred to the target database for fine-tuning. To take advantage of temporal information, we have utilized the Long Short Term Memory (LSTM) network to consider the sequences of CNN actuations explicitly. CNNs like VGG-16 and AlexNet, which are pre-trained on ImageNet, can be used as a feature extractor.

Spatial Feature Extraction

In order to make full use of static databases, we have used VGG-16 architecture for dimensional feature extractions. The VGG-16 is the deep convolutional network with up to sixteen layers (thirteen convolutional layers and three fully connected layers). This network takes an input image size of 224 * 224 pixels, with a convolutional kernel size of 3 * 3 and max-pooling with 2 * 2 windows. We used the pre-trained VGG-Face [91] architecture to initialize the network parameters that are trained on a massive facial dataset of 2.6 million images. We assume that databases that are captured in controlled and uncontrolled environmental conditions with posed as well as spontaneous expression are involved, and we use the transfer learning strategy to transfer the "information" learned by the VGG-model to our new target dataset of neurocenter residents suffering from brain injuries for emotional cues identification. Transfer learning can be used to avoid overfitting in the training of our network, considering the TBI database is too limited in terms of identities of subject to train a generalized network.

The LSTM for Temporal Information Extraction

In general, CNNs deal with images that are isolated. However, in our case, we have used sequences of images as well, thus preserving the temporal information. LSTM models are capable of absorbing this dynamic sequential information. The LSTM modules can determine long-range temporal correlations from the input sequences by using memory cells, which can hold and release information.

As illustrated in figure E.5, the LSTM states are controlled by three gates associated with forget (f), input (i), and output (o) states. These gates regulate the flow of information through the model by using point-wise multiplications and sigmoid functions σ, which bind the information flow between zero and one by the set of mathematical equations as explained in [2, 3]. The datasets used to train the CNN were chosen from the benchmark datasets publicly available or made available to the research community, and they are described in section 2.1.

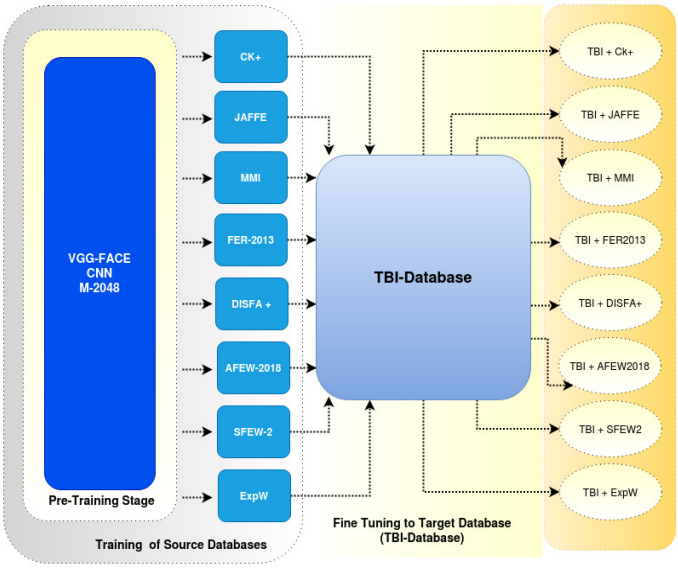


Fig. E.6: Databases explored for Transfer learning

4.3 Transfer Learning Mechanism

In the current research project, we have to deal with limited labeled and identity data from people that suffered a traumatic brain injury. However, learning processes in deep neural networks need lots of labeled training data. Gathering training data and labeling it is very difficult and time-consuming work. So, for gaining more accurate results, we make use of new techniques such as transfer learning.

Transfer learning is a powerful technique which adapts knowledge from some related auxiliary well-labeled source domains. Considering the benefit of transfer learning, we can use labeled data that was gathered with healthy subjects to optimize target

data. In general, transfer learning methods categorize into two groups: domain-invariant feature learning and classifier adaptation. In this paper, we applied an in-depth transfer learning approach to unify the knowledge transfer and deep feature learning.

Since the input of our architecture are image frames and image sequences, so we had implemented the learning of features in two ways; firstly by the use of only static images and transferring the knowledge to the TBI datasets; secondly exploiting the dynamic features of video data, as the variations between image sequences encode additional advantageous information for the classification of emotional signals.

Similar to the work in [53, 55, 56], we employed the VGG-16 model to initialize the network parameters and learn the features from eight public databases. Since the bottom layers of CNNs learn more generic features and top layers acquire more sophisticated and data specific information [92], so we reserved only the convolutional and max-pooling layers and discarded the pre-trained last three fully-connected layers. We removed fully-connected layers as they do not hold spatial information, which is essential for the capture of motion signals in the subsequent LSTM model. Therefore, the last pooling layers of the CNN framework is linked directly to the LSTM to investigate the temporal characteristics across coherent images.

5 Experimental Results

In this section, we evaluate the performance of our proposed model in two ways: First, by the domain transfer learning of static as well as dynamic databases to our target TBI database. Second, by evaluating the emotional cues learned and transferred from controlled and uncontrolled environmental conditions to the TBI datasets. A static dataset refers to the image frames, whereas dynamic relates to the sequences of images or video sequences.

5.1 Experimental Results Evaluation for Static Datasets

The facial images are resized to 224 * 244 pixels according to the network-input parameter. Peak expression frame is used for training of the network for CK+, MMI, DISFA+ datasets. JAFFE, FER2013, SFEW, ExpW have mostly one to four images per expression. Video datasets are first converted into 30 frames per second by an open-source video converter, and then the peak expression image is selected. Data is distributed 80% for training and 20% for testing purposes. The network is trained with a learning rate of 0.0001, and batch normalization is applied to normalize the input layer.

Figures E.7 illustrates the performance of our models trained on eight different datasets. We can identify that recognition performance of contempt is not good as compared to other expressions through the confusion matrix in Fig. E.7(a). Besides, we can determine that fear and disgust emotion expressions are less accurate, as demonstrated by the confusion matrix in Fig.E.7(c). However confusion matrix of datasets captured in controlled environment Fig.E.7 (a)(b)(c)(d) have much higher performance than of in-the-wild setting databases as evident in the Fig.E.7(e)(f)(g)(h) for emotional categories. The overall accuracies of our proposed network are compared with other state-of-the-art methods, as seen in table E.5, and it is observed that

our model has performed competitively well.

Table E.5: Performance evaluation of our (VGG-FineTuned) model for emotional categories for static datasets with other results in the literature in terms of average accuracy.

Group	Method	Training Parameters	Accuracy (%)
CK+	Liu et al. [57]	8 folds	97.1
	Zhang et al. [29]	10 folds	**98.9**
	Our	10 folds	98.6 $\pm$0.59
JAFFE	**Liu et al. [39]**	LOSO	**91.8**
	Our	10 folds	89.46 $\pm$1.75
MMI	Liu et al. [57]	10 folds	78.53
	Li et al. [58]	5 folds	78.46
	Our	10 folds	**79.06 $\pm$0.88**
DISFA+	**Our**	5 folds	**77.15 $\pm$4.92**
FER 2013	Zhang et al. [29]	Training 28,709	Test 75.1
	Tang et al. [70]	Validation 3,589	Test 71.2
	Kim et al. [59]	Test 3,589	Test 73.73
	Our	Training 35887 Validation 7178	Val **78.19 $\pm$2.47**
SFEW	Li et al. [58]	Training 958 , Validation 436 , Test 372	Val 54.19 (47.97)
	Meng et al. [60]		Val 50.98 (42.57) Test: 54.30 (44.77)
	Yu et al. [61]		**Val 55.96 (47.31) Test 61.29 (51.27)**
	Our		Val 55.75 $\pm$2.74

5.2 Dynamic Database

The temporal information exploration is analyzed on four publicly available datasets, namely CK+, MMI, DISFA+, and AFEW. The performance of fine-tuned VGG-face model is compared with state-of-the-art methods in Table E.6. It is clearly observed that in the case of the DISFA+ dataset, our network has produced better results. Similarly, our network has surpassed the state-of-the-art methods in case of AFEW dataset, when tested on the validation set. For CK+ and MMI datasets, our fine-tuned model produced decent and competitive results. The confusion matrices to represent the accuracies of seven emotional categories are illustrated in the Fig. E.8. Fig. E.9 represents the performance of our architecture employed to static and dynamic datasets. It is evident that temporal information has increased the performance of the network.

Table E.6: Performance evaluation of our (VGG-finetuned) model for emotional categories for Dynamic datasets with other results in the literature in terms of average accuracy.

Group	Method	Training Parameters	Accuracy (%)
CK+	Zhao et al. [62]	Training: 7 to last frame Test: last frame; 10 folds	99.3
	Yu et al. [63]	Training: 7 to last frames Test: peak expression; 10 folds	**99.6**
	Kim et al. [64]	All frames used in training and testing; 10 folds	97.93
	Zhang et al. [65]	All frames used in training and testing; 10 folds	98.50
	Kuo et al. [66]	9 frames for training and testing; 10 folds	98.47
	Our	10 folds	98.92 $\pm$0.32
MMI	Kim et al. [64]	LOSO	81.53
	Zhnag et al. [65]	All frames for training and testing; 10 folds	81.18
	Sun et al. [67]	10 folds	**91.46**
	Our	10 folds	85.89 $\pm$1.52
DISFA+	Zhang et al. [65]	All frames for training and testing; 10 folds	93%
	Our	10 folds	**94.09 $\pm$0.77**
AFEW	Otberdout et al. [68]	Training 773, Validation 373, Test 593 videos	Val 46.32 Test 49.59
	fan et al. [69]		45.43 on Val
	Fan et al. [69]		59.02 on Test
	Our		Val **50.17 $\pm$1.68**

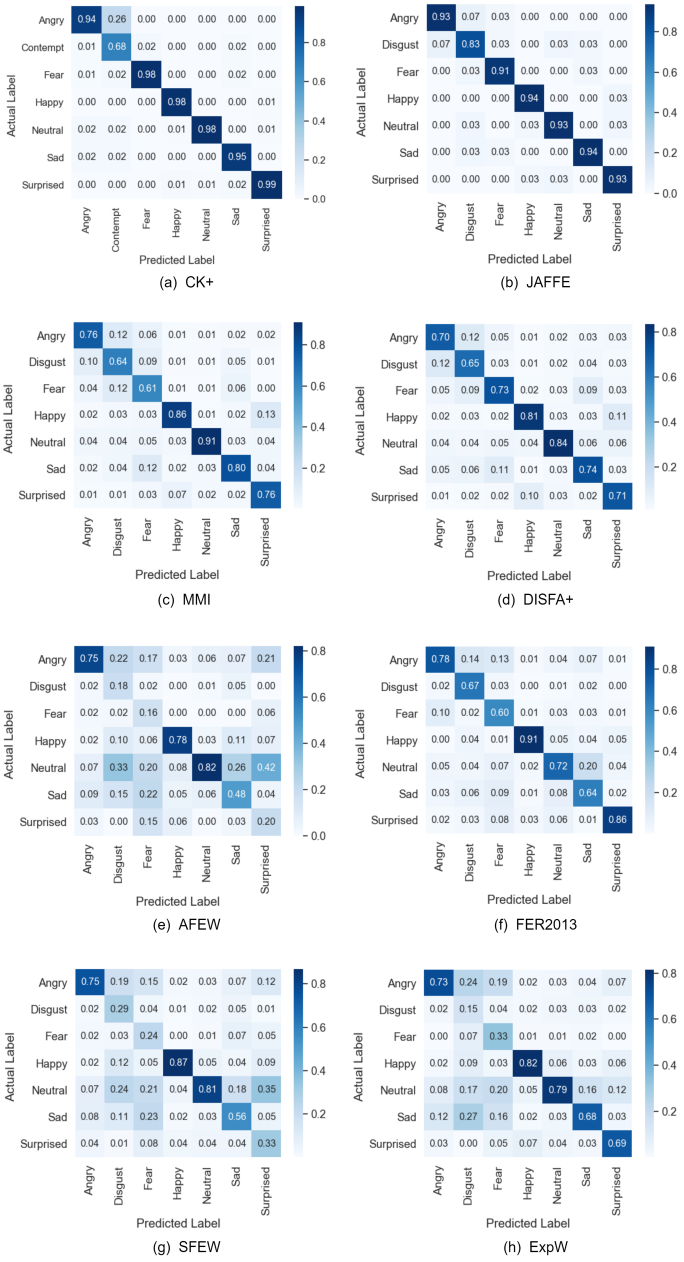


Fig. E.7: Performance visualization of the models trained on eight source databases using image frames.

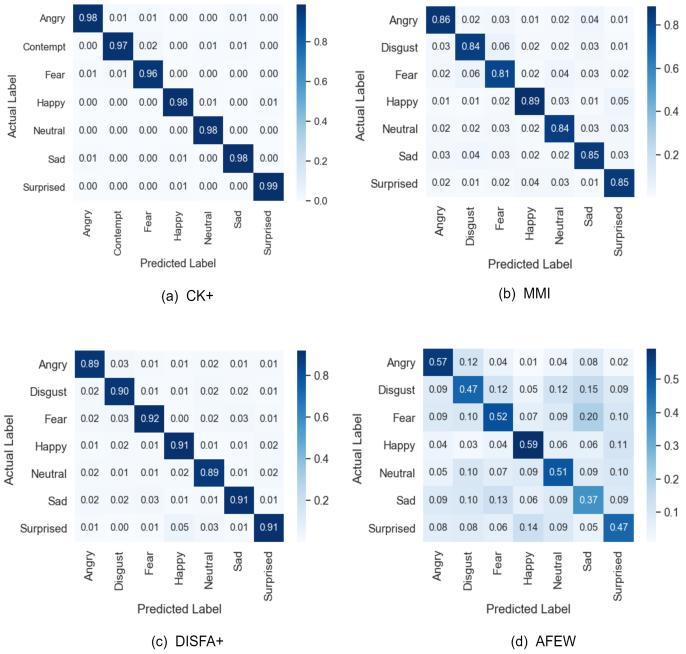


Fig. E.8: Performance visualization of models trained on four source databases using sequences of images.

5.3 Contribution in Emotion Recognition

In the second stage, the target 'TBI datasets' are fine-tuned with pre-trained and tuned VGG-face model with the above/mentioned publicaly available datasets, in both static and dynamic formats. Despite the challenges of less-expressing and limited-identity datasets, fine-tuned model exhibited the comparable results. In our experimentation, we executed single-source-single-target transfer learning, that is individual source dataset features are transferred to TBI dataset and then emotions are classified. Our network learned the facial features related to the specific emotional category of healthy people and explored those characteristics into facial features of TBI-datasets.

5.4 Evaluation Metric

We evaluated the performance of our framework using evaluation matrices to fully understand the model efficacy. Confusion matrices, precision, recall, Area under curve (AUC), and the average accuracy present the performance of our model to recognize subtle emotional changes. We calculated multi-class confusion matrices for both static and dynamic datasets as well as before and after fine tuning to the target datasets as shown in Fig. E.7, Fig. E.8, Fig. E.12, and Fig E.11.

To understand the strength of each dataset for a particular expression category, we employed precision and recall matrices as illustrated in Table E.7 and Table E.8. Re-

sults demonstrate that dataset captured in the wild such as AFEW, SFEW, and ExpW have lesser accuracy for disgust, fear and surprise expressions. However, FER2013 performed quite well for the same expressions. We identified that mis-classification of these emotions could be due to a lower number of such expressions in the datasets under analysis. A trend of increase in accuracy for each emotional class is witnessed with an increase in number of frames. Overall, the precision-recall matrices work in relationship; precision indicate the ability of model to determine only relevant data points whereas recalls verify that determined data points are actually relevant.

As given in the equations, we determined accuracy, precision and recall:

$$Accuracy = \frac{TP + TF}{TP + TF + FP + FN} \tag{E.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{E.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{E.3}$$

where TP, TN, FP, and FN are the overall true positive, true negative, false positive, and false negative of all the classes in the confusion matrix. In other words, the overall accuracy was the sum of off-diagonal elements divided by all the elements in the multi-class confusion matrix. Table E.5 demonstrates the performance of our model on static source datasets. It is evident that our model has performed better with accuracy of 79.06% and 78.19%, surpassing 78.53% and 73.73% on MMI and FER2013-validation dataset respectively. On the rest of datasets, our model competed state-of-art-methods while measuring frame-based accuracies. On contrast, our model with sequential information has fared well surpassing recognition accuracies by 94.09% and 50.17% on the DISFA and the AFEW datasets respectively as presented in Table E.6. We have used average accuracy metric due to imbalanced emotional data as mentioned in Table E.2. For additional performance measure, statistical significance of emotional recognition is verified by t-test conducted on all datasets.

Fig. E.9 provides the illustration of overall performance of the network exploiting the static as well as temporal information from the various source datasets. It also demonstrates the performance of the network on the target TBI challenging dataset after fine-tuning with source datasets. It is evident from the results that use of temporal information have enhanced the accuracy as it is evidenced through AUC metrics in Fig. E.10, where static and temporal information are considered in the model training. In addition, fine-tuning with various source datasets exhibited that performance is dependent on two factors; One is more training data facilitates better in transfer of features and secondly, features related to negative emotions are learnt better from the datasets captured in controlled settings. It is seen from the confusion matrices in Fig. E.11 and Fig. E.12, that accuracy of emotional expressions of anger, contempt/disgust and fear is better when fine-tuned with CK+, MMI, and DISFA+ as compared to AFEW.

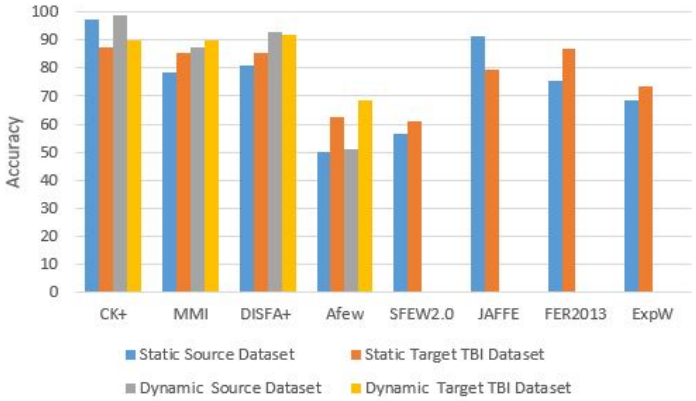


Fig. E.9: Source vs Traget datasets accuracy comparison: Illustration provides the performance of the network when static and temporal information from both source and target datasets is utilized. For CK+, MMI, DISFA+ and AFEW datasets we have used both static and dynamic information, whereas for SFEW, JAFFE, FER2013 and ExpW static information is explored.

Table E.7: Precision matrix for each expression class for source datasets

	Precision							
Expressions	**CK+**	**JAFFE**	**MMI**	**DISFA**	**AFEW**	**FER2013**	**SFEW**	**ExpW**
Anger	96.67	90.01	78.87	69.72	77.55	79.31	77.78	74.49
Contempt	88.89	-	-	-	-	-	-	-
Disgust	-	82.76	67.56	64.814	18.5	68.68	29.59	66.03
Fear	96	93.75	66.79	72.91	16.32	61.45	24.51	33.67
Happy	98.56	96.77	83.95	81.03	83.83	92.89	88.64	83
Neutral	98.76	90	85.13	83.54	83.67	73.77	82.89	81.25
Sad	94.64	93.55	77.82	73.63	48.45	64.57	57.60	71.38
Surprise	96.75	93.32	77.42	71.42	20.40	87.60	32.85	70.47

6 Insights on Emotion Recognition in the Rehabilitation of TBI Patients

The rehabilitation phase usually requires four steps [93]. First, the impairment type and its severity must be tested. Second, the therapist set rehabilitation goals. Third, the rehabilitation intervention takes place. Finally, following the intervention, the patient has to be re-evaluated, allowing to adjust the objectives. Robots have the potential to assist and promote rehabilitation procedures. They can be used to measure performance prior, during and after an intervention as well as systematically and continuously suggest treatment strategies based on this input and the severity of the disability. The intervention of the Pepper robot integrated with customized emotion recognition module assisted the rehabilitation process for the TBI patients in the four phases, as mentioned earlier. In our field study, first, we studied the impairment severity of each patient, and pre-set targets were defined and tested during and after the intervention of the pepper robot. In our case, we distribute the pepper robot as-

Table E.8: Recall matrix for each expression class for source datasets

	Recall							
Expressions	**CK+**	**JAFFE**	**MMI**	**DISFA**	**AFEW**	**FER2013**	**SFEW**	**ExpW**
Anger	94.56	93.10	75.95	71.53	50.67	66.95	57.48	62.39
Contempt	86.48	-	-	-	-	-	-	-
Disgust	-	82.66	63.86	66.96	66.67	90.66	70.31	82.5
Fear	94.56	88.91	61.24	5574	64.21	7.64	61.53	84.22
Happy	97.84	93.75	85.76	92.23	70.33	83.63	72.13	79.04
Neutral	98.43	93.01	90.98	95.68	37.61	65.17	43.38	55.03
Sad	96.36	91.31	79.92	23.32	44.34	70.01	50.96	60
Surprise	98.77	93.33	75.98	17.74	48.75	82.07	59.25	79.57

sistance in two categories; robot as a monitoring agent and as a Feedback agent for both patients and therapists.

6.1 Pepper robot as a Monitoring Agent

The intervention with the Pepper robot has been designed to in relation to the three scenarios used during the data collection: cognitive, physical and social interaction rehabilitation. The first phase involves the determination of the impairment level for each scenario. It is determined with the set the of protocols and disability condition as mentioned in the table E.1. In our pilot study we determine the emotional expressions before, during and after each rehabilitation strategy. Before the deployment of the pepper interventions, the data collected was extremely beneficial for the clinician and therapist to evaluate how cognitive learning, physical movement and social interaction patterns can be affected with changes in the expressions. For example, in cognitive rehabilitation tasks, subjects tend to make mistakes when there are more negative emotional expressions. Therefore, in such a case, the performance of the subject declines. Similarly, patients are hesitant to involve or sometimes resist to indulge in physiotherapy tasks when they are tired or exhibit negative emotions. In such a scenario, the therapist failed to achieve targets, set for the rehabilitation exercise. In case of social interaction activity it is observed that passive stimulus is required to enhance social interaction, where subjects hardly communicate with other subjects or passively communicate with therapists.

Monitoring Negative Emotional Reactions

Research conducted in [2] illustrates that to achieve the best results, it is essential to determine the emotional states of the patients prior to conducting an rehabilitation exercise. This would have a large impact on an effective rehabilitation as therapists could save time and effort and eventually adapt rehabilitation strategies based on the emotional conditions of the patients. For this purpose, the Pepper robot intervention facilitates the staff members and therapists to determine the emotional states before, during and after the rehabilitation tasks. In addition, Pepper generates reactions according to an individual patient's emotional state to assist in achieving the targets set for the rehabilitation exercise.

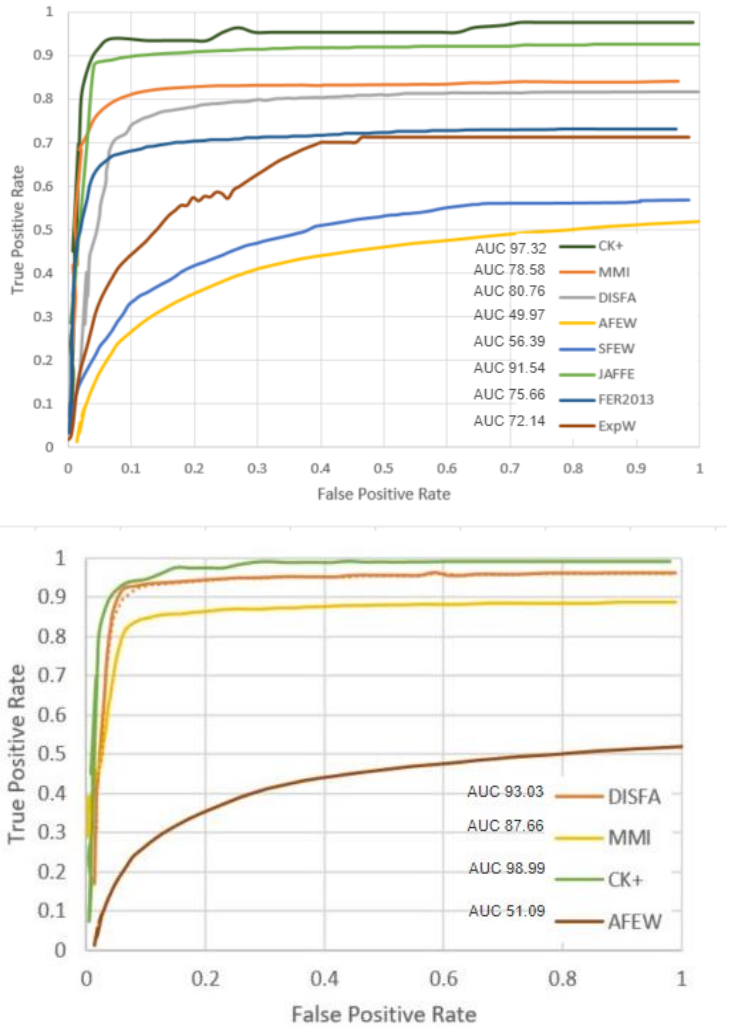


Fig. E.10: ROC curves for emotion recognition through frame-based and sequence of images based information.

Handling Negative Reactions

In our pilot study, Pepper uses audio, visual and gesture output to handle negative emotional reactions generated by the patients during rehabilitation tasks. The robotic intervention impacted positively on physical rehabilitation but negatively on cognitive activity. In case of physical rehabilitation, patients were motivated to execute more repetitions of tasks. However, patients find the Pepper robot intervention distracting during the cognitive tasks. This is due to the fact that during cognitive activity, Pep-

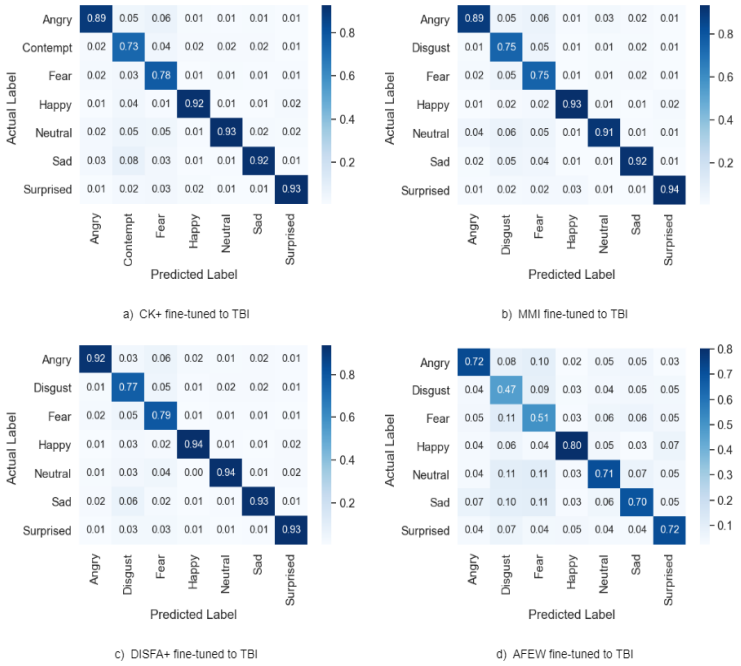


Fig. E.11: Performance visualization of models fine-tuned to the target TBI database exploiting temporal information from the image sequences.

per identified their focused-emotional reactions as negative expressions and reacted accordingly. We implemented a Wizard-of-Oz (WoZ) functionality to recognise behavioral traits in humans to equip the Pepper robot with intellectual cognitive abilities in decision making as well as in creating good relationships with its human user. The WoZ feature aids the therapists to achieve the rehabilitation targets during cognitive task execution and also supports building a reliable relationship between robot and human user.

Performance Monitoring

Pepper records each rehabilitation session and generates a pool of expressions over time as illustrated in the Fig. E.13. The pool of expressions determines the accuracy over the rate of change of expressions from positive to negative and vice versa. In our pilot study, we analyzed subjects exhibit positive expressions while accurately execution of the physical and cognitive tasks. In case of cognitive assessment, pool of expressions are also compared with the results of Android application "Luminosity" that keeps the track of accuracy over the entire session as well as for repetitive tasks for each individual subject. These results also confirmed the exhibition of positive expressions with accuracy of tasks accomplished. During physiotherapy, Pepper acted as a "motivator" that resulted in more repetitions of physical activity during a session for the majority of the patients. The number of robotic reactions in response to positive emotional expressions is directly proportional to the the number of repeti-

tions executed in a given session. For instance, in our case study when pepper robot is placed with the subject, number of reps for physiotherapy were increased significantly so the Pepper reactions to acknowledge the effort and motivate the subject. Fig. E.14 illustrates the Pepper robot interaction with a subject while executing the physiotherapy activity.

6.2 Pepper robot as a Feedback agent

Conventional evaluations involve one-on-one consultations with a therapist. Employing Pepper supports this approach with an objective evaluation of motor and cognitive functions utilizing data obtained during rehabilitation sessions, thus, allowing for accurate, effective, and automated evaluation of motor and cognitive abilities independent of human biases. In addition, audio, visual and gesture output of the Pepper robot during the activity, can provide information about patient-activity-engagement and attention time-span. Attention span of TBI patients is generally low, however, with robotic intervention this issue can be minimized using emotional expression information, where a therapist need to modify the activity to maintain the interest of the subject. This feedback with robotic output and pool of expressions enable the therapists to modify the treatment according to patient involvement and performance.

6.3 Challenges and Limitation

We will discuss challenges and limitation related to emotion analysis system and robotic platform and rehabilitation strategies involved as follows.

Comparing the facial expression recognition accuracy with others work is quite challenging as different researchers adopt different databases with varying pre-processing techniques and training techniques. Despite we do performance comparisons with methods explored and average accuracy achieved. we need to consider the balanced and imbalanced data within expression categories for metrics evaluation. Table E.7 and Table E.8 presents the performance variance of network with varying data classes. Therefore, it is necessary to apply relevant evaluation matrices for system performance analysis.

Although the treatment for rehabilitation through robotic interventions have been proven to be beneficial, in most facilities they are not yet part of standard care. This is mainly due to the fact that most studies have been carried out with non-mass-developed robotic devices, even though commercially produced social rehabilitation robots are becoming popular, but their costing rise significantly. Along with the need to include more people with clinical rehabilitation substantial attempts are now being made to create and implement low-cost tools that mitigate direct therapist oversight. In the neuro centers, a big obstacle for introducing robot-assisted therapy is that the patient must be able to adhere with the recommended procedure. The patient adherence to recommended treatments in therapy is correlated with both decreased compliance and improved treatment outcome. However, lack of desire to do the workouts is one of the key reasons for the inability to adhere. Introduction of more engaging interface such as utilization of the Pepper robot display, synchronized with robotic gestures and audio framework could contribute towards persistent motivation. In addition, where patients impairments are severe, the system can respond by allowing

the therapist taking control over the robotic intervention to modify the treatment.

7 Conclusion

In this work we have contributed in two phases, first towards the development of emotion recognition algorithm for TBI patients and second the deployment of the robotic framework for rehabilitation of the TBI patients through the implementation emotion recognition model. For emotion recognition, we have introduced a deep learning framework that is trained to learn the facial features from the datasets acquired in controlled and uncontrolled environment to address two major issues in automatic facial expression recognition. The first problem that we address in this work is non uniform display of human facial expressions. For instance, in case of TBI patients where facial expressions are variant due to artifacts caused by impairment severity. Employing CNN and CNN-LSTM algorithm we transferred static and dynamic facial characteristics related to each expressions to TBI patients database having limited identities. one one hand, our methods have achieved the state-of-the-art performances on specific datasets in both frame-based (static) and sequence of frames based (dynamic) emotional recognition. Our model has improved the accuracy on various datasets,for instance 78.53% to 79.06% on MMI and 73.73% to 78.19% on FER2013 database in static analysis. Similarly, use of temporal information had enabled the network to exhibit state-of-art performance on DISFA and the AFEW with 94.09% and 50.17% accuracy results respectively as presented in Table E.6.

On the other hand, our experimental studies reveal that certain facial expressions like anger, fear contempt/disgust, sad and surprise are learnt better from the databases that posses features with frontal faces such as CK+, MMI and JAFFE. Whereas facial features related to neutral, happy expressions have exhibited constant learning pattern in both controlled and in-the-wild environmental conditions. However, large databases in-the-wild like FER2013 and ExpW have produced better results than smaller databases. In addition, posed facial expressions in lab or controlled environment, are impure and inconsistent that cause significant degrading in performance of facial expression algorithms in the real world settings. In this work, we train our CNN-LSTM model to transfer facial features in-the-wild settings to the TBI database having pure expressions that were carefully annotated by the experts and clinical staff members, increasing the FER accuracy on the TBI images. Our experimental findings indicate that the proposed FER algorithm achieves equal or even better performance than state-of-the-art methods.

The second major contribution is the use of robotic technology to transform the recovery from a one-on-one comprehensive care of human beings in specialized institutions to a technologically-driven, centrally monitored and controlled environment. Provided the elevated costs associated with long-term recovery and the challenge in maintaining adequate duration and severity of impairment treatment rehabilitation programs, cost-effective deployment of robotic rehabilitation is firmly supported. Implementing emotion understanding through the Pepper robot empowers clinicians to deliver more productive recovery interventions and enable patients to access care more efficiently.

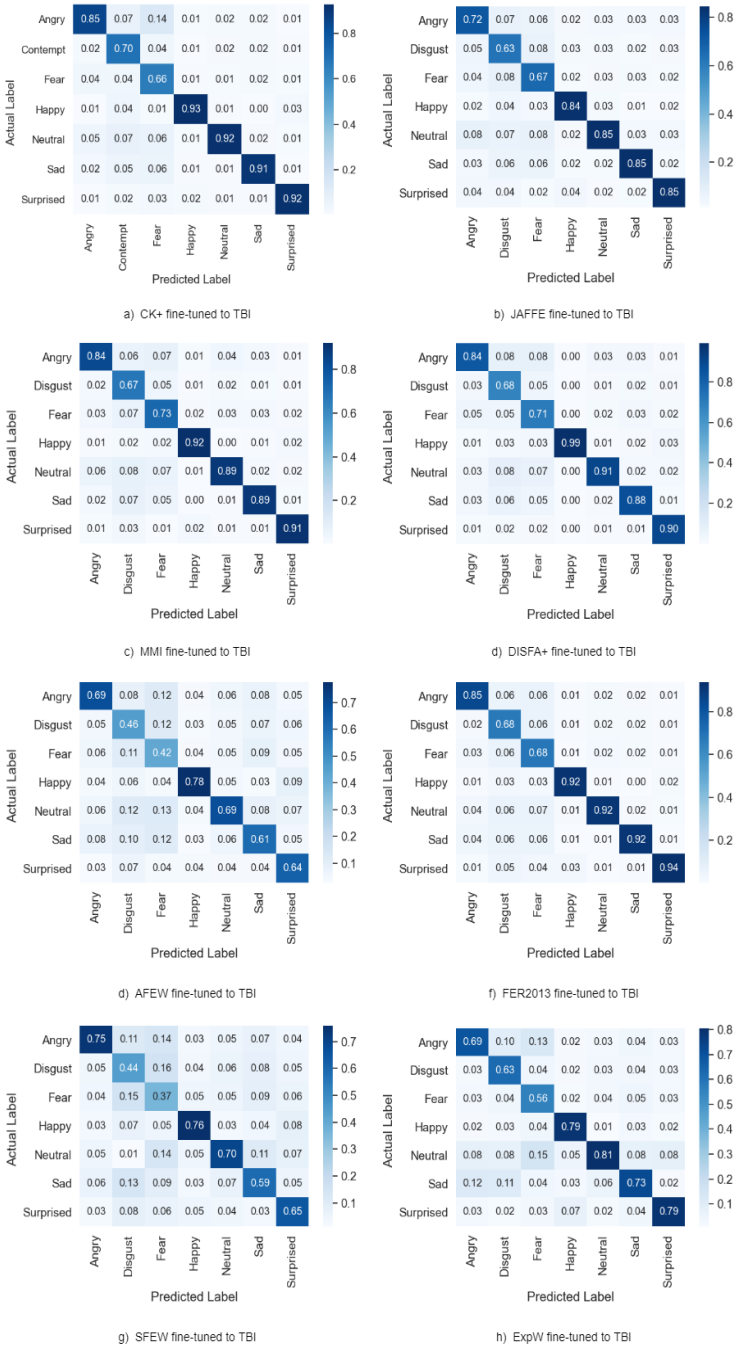


Fig. E.12: Performance visualization of models fine-tuned to the target TBI database using image frames.

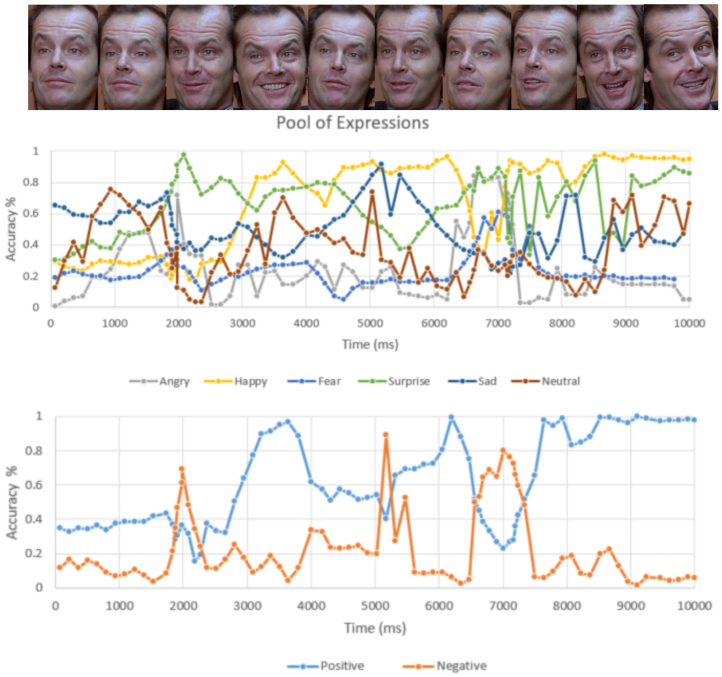


Fig. E.13: Visualization of pool of expression in timely order. Video sample of maximum 10 second is taken from AFEW dataset and every 25th frame per second is displayed

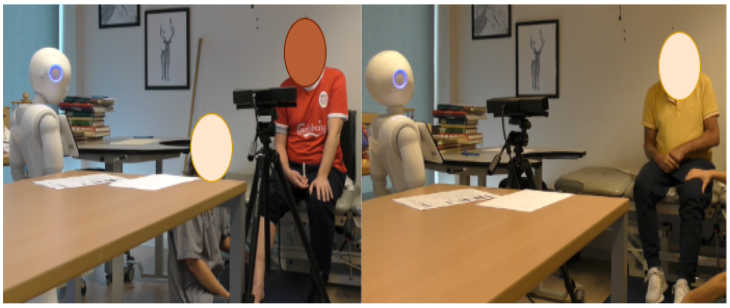


Fig. E.14: Visualization of the Pepper robot interaction with the subjects during physical rehabilitation activity. Identities are not covered due to privacy issues.

References

[1] K. Rodil, M. Rehm, and A. L. Krummheuer, "Co-designing social robots with cognitively impaired citizens," in *The 10th Nordic Conference on Human-Computer InteractionNordic Conference on Human-Computer Interaction*. Association for Computing Machinery, 2018.

[2] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 369–389.

[3] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[4] D. Calvaresi, D. Cesarini, P. Sernani, M. Marinoni, A. F. Dragoni, and A. Sturm, "Exploring the ambient assisted living domain: a systematic review," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 2, pp. 239–257, 2017.

[5] R. Maskeliūnas, R. Damaševičius, and S. Segal, "A review of internet of things technologies for ambient assisted living environments," *Future Internet*, vol. 11, no. 12, p. 259, 2019.

[6] V. H. C. d. Albuquerque, R. Damaševičius, N. M. Garcia, P. R. Pinheiro *et al.*, "Brain computer interface systems for neurorobotics: methods and applications," 2017.

[7] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer handbook of robotics*. Springer, 2016, pp. 1657–1684.

[8] J. Burgner-Kahrs, D. C. Rucker, and H. Choset, "Continuum robots for medical applications: A survey," *IEEE Transactions on Robotics*, vol. 31, no. 6, pp. 1261–1280, 2015.

[9] H. Robinson, B. MacDonald, and E. Broadbent, "The role of healthcare robots for older people at home: A review," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 575–591, 2014.

[10] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. De Witte, "Socially assistive robots in elderly care: A systematic review into effects and effectiveness," *Journal of the American Medical Directors Association*, vol. 13, no. 2, pp. 114–120, 2012.

[11] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, p. eaao6760, 2018.

[12] J.-J. Cabibihan, H. Javed, M. Ang, and S. M. Aljunied, "Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism," *International journal of social robotics*, vol. 5, no. 4, pp. 593–618, 2013.

[13] R. Adolphs, "Neural systems for recognizing emotion," *Current opinion in neurobiology*, vol. 12, no. 2, pp. 169–177, 2002.

[14] R. M. Müri, "Cortical control of facial expression," *Journal of comparative neurology*, vol. 524, no. 8, pp. 1578–1585, 2016.

[15] M. Balconi, "Neuropsychology of facial expressions. the role of consciousness in processing emotional faces," *Neuropsychological Trends*, vol. 11, pp. 19–40, 2012.

[16] J. Šalkevicius, R. Damaševičius, R. Maskeliunas, and I. Laukienė, "Anxiety level recognition for virtual reality therapy system using physiological signals," *Electronics*, vol. 8, no. 9, p. 1039, 2019.

[17] N. M. Krishna, K. Sekaran, A. V. N. Vamsi, G. P. Ghantasala, P. Chandana, S. Kadry, T. Blažauskas, and R. Damaševičius, "An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using eeg signals," *IEEE Access*, vol. 7, pp. 77 905–77 914, 2019.

[18] C. M. A. Ilyas, V. Schmuck, M. A. Haque, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Teaching pepper robot to recognize emotions of traumatic brain injured patients using deep neural networks," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[19] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[20] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

[22] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. Paris, France, 2010, p. 65.

[23] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The japanese female facial expression (jaffe) database," in *Proceedings of third international conference on automatic face and gesture recognition*, 1998, pp. 14–16.

[24] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[25] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–8.

[26] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, vol. 2, p. 1, 2011.

[27] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 524–528.

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[30] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[31] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 211–216.

[32] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[33] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[34] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[35] Y. Stern, "Cognitive reserve," *Neuropsychologia*, vol. 47, no. 10, pp. 2015–2028, 2009.

[36] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014. [Online]. Available: https://www.pnas.org/content/111/15/E1454

[37] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2106–2112.

[38] W. V. Friesen, P. Ekman *et al.*, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983.

[39] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[40] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[41] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*. Cham: Springer International Publishing, 2017, pp. 151–162.

[42] J. Wan, S. Escalera, G. Anbarjafari, H. Jair Escalante, X. Baró, I. Guyon, M. Madadi, J. Allik, J. Gorbova, C. Lin *et al.*, "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3189–3197.

[43] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49–61, 2018.

[44] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "Pca-based dictionary building for accurate facial expression recognition via sparse representation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1082–1092, 2014.

[45] S. Berretti, B. Ben Amor, M. Daoudi, and A. del Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, p. 1021, Jun 2011. [Online]. Available: https://doi.org/10.1007/s00371-011-0611-x

[46] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," *Sensors*, vol. 11, no. 10, pp. 9573–9588, 2011.

[47] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803 – 816, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608001844

[48] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[49] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using hog–ebgm," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537 – 1543, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508001104

[50] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on multimodal interaction*. ACM, 2014, pp. 494–501.

[51] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 451–458.

[52] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 459–466.

[53] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443–449.

[54] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "Holonet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 472–478.

[55] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2017.

[56] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.

[57] X. Liu, B. Vijaya Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–29.

[58] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[59] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 48–57.

[60] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 558–565.

[61] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[62] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016, pp. 425–442.

[63] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer*, vol. 34, no. 12, pp. 1691–1699, 2018.

[64] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," *arXiv preprint arXiv:1703.07140*, 2017.

[65] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[66] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2121–2129.

[67] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, vol. 119, pp. 49–61, 2019.

[68] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Automatic analysis of facial expressions based on deep covariance trajectories," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[69] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 445–450.

[70] Y. Tang, "Deep learning using support vector machines," *CoRR*, vol. abs/1306.0239, 2013. [Online]. Available: http://arxiv.org/abs/1306.0239

[71] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[72] K. Kulkarni, C. Corneanu, I. Ofodile, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of facial displays of unfelt emotions," *IEEE transactions on affective computing*, 2018.

[73] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baró, H. Demirel *et al.*, "Dominant and complementary emotion recognition from still images of faces," *IEEE Access*, vol. 6, pp. 26 391–26 403, 2018.

[74] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[75] A. Barman, A. Chatterjee, and R. Bhide, "Cognitive impairment and rehabilitation strategies after traumatic brain injury," *Indian Journal of Psychological Medicine*, vol. 38, no. 3, pp. 172–181, May-Jun 2016.

[76] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, pp. 5–pp.

[77] A. Shapi'i, M. Zin, N. Azan, and A. M. Elaklouk, "A game system for cognitive rehabilitation," *BioMed research international*, vol. 2015, 2015.

[78] J. C. Perry, J. Andureu, F. I. Cavallaro, J. Veneman, S. Carmien, and T. Keller, "Effective game use in neurorehabilitation: user-centered perspectives," in *Handbook of research on improving learning and motivation through educational games: multidisciplinary approaches*. IGI Global, 2011, pp. 683–725.

[79] A. M. Elaklouk, N. A. M. Zin, and A. Shapii, "Investigating therapists' intention to use serious games for acquired brain injury cognitive rehabilitation," *Journal of*

King Saud University-Computer and Information Sciences, vol. 27, no. 2, pp. 160–169, 2015.

[80] L. Rapple, "Lotsa helping hands," *FOCUS: Journal for Respiratory Care and Sleep Medicine*, p. 36, Jul 1, 2008.

[81] T. Tsaousides and W. A. Gordon, "Cognitive rehabilitation following traumatic brain injury: assessment to treatment," *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 76, no. 2, pp. 173–181, 2009. [Online]. Available: http://dx.doi.org/10.1002/msj.20099

[82] L. Rees, S. Marshall, C. Hartridge, D. Mackie, and M. W. F. T. E. Group, "Cognitive interventions post acquired brain injury," *Brain Injury*, vol. 21, no. 2, pp. 161–200, 2007. [Online]. Available: https://doi.org/10.1080/02699050701201813

[83] K. McKenna, D. M. Cooke, J. Fleming, A. Jefferson, and S. Ogden, "The incidence of visual perceptual impairment in patients with severe traumatic brain injury," *Brain Injury*, vol. 20, no. 5, pp. 507–518, 2006. [Online]. Available: https://doi.org/10.1080/02699050600664368

[84] M. Sutton, "Apps to aid aphasia," *ASHA Leader*, vol. 17, no. 7, p. 32, Jun 1, 2012. [Online]. Available: https://search.proquest.com/docview/1022993653

[85] J. A. Hugentobler, M. Vegh, B. Janiszewski, and C. Quatman-Yates, "Physical therapy intervention strategies for patients with prolonged mild traumatic brain injury symptoms: a case series," *International journal of sports physical therapy*, vol. 10, no. 5, p. 676, 2015.

[86] M. Jones and P. Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, no. 14, p. 2, 2003.

[87] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[88] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[89] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[90] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4295–4304.

[91] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[92] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[93] P. Langhorne, J. Bernhardt, and G. Kwakkel, "Stroke rehabilitation," *The Lancet*, vol. 377, no. 9778, pp. 1693–1702, 2011.

Paper F

Teaching Pepper Robot to Recognize Emotions of Traumatic Brain Injured Patients Using Deep Neural Networks

Chaudhary Muhammad Aqdus Ilyas, Viktor Schmuck, Muhammad Ahsanul Haque, Kamal Nasrollahi, Matthias Rehm, and Thomas B. Moeslund

The paper has been published in the
Proceedings of 2019 28th IEEE International Conference on Robt and Human

Interactive Communication (RO-MAN 2019),
doi:10.1109/RO-MAN46459.2019.8956445
2019.

The layout has been revised.

Abstract

Social signal extraction from the facial analysis is a popular research area in human-robot interaction. However, recognition of emotional signals from Traumatic Brain Injured (TBI) patients with the help of robots and non-intrusive sensors is yet to be explored. Existing robots have limited abilities to automatically identify human emotions and respond accordingly. Their interaction with TBI patients could be even more challenging and complex due to unique, unusual and diverse ways of expressing their emotions. To tackle the disparity in a TBI patient's Facial Expressions (FEs), a specialized deep-trained model for automatic detection of TBI patients' emotions and FE (TBI-FER model) is designed, for robot-assisted rehabilitation activities. In addition, the Pepper robot's built-in model for FE is investigated on TBI patients as well as on healthy people. Variance in their emotional expressions is determined by comparative studies. It is observed that the customized trained system is highly essential for the deployment of Pepper robot as a Socially Assistive Robot (SAR).

1 INTRODUCTION

Researchers have conducted extensive investigation into human-focused robotic technologies, designed to achieve real time and close to human-like human-robot interactions [1]. However, existing robotic technologies that facilitate robots in human emotions recognition have limitations [1] and require more intelligent platforms and software to communicate and respond naturally with people [2]. Recently robots have been developed to collaborate with doctors, physicians or physiotherapists. In the health care sector these robots are tailored-made, particularly Socially Assistive Robots (SAR), to provide assistance and improvement in a wide range of medical applications such as robot-assisted therapies [3, 4], complex-surgical operations [5, 6], or for social engagement with people with special needs like children with autism spectrum disorder (ASD) [7–10]. Machine learning, especially deep learning, approaches have enabled these robots to automatically identify and react intelligently to subject emotional states. These smart machines require techniques that can accurately and robustly recognize human emotional clues from uncontrolled and natural environmental conditions [11].

A typical robot for health monitoring and improvement needs to receive audio, video or proximity information from its sensors. This information is then processed based on the algorithm that interpret the information into meaningful signals. This is followed with robot action or response for the desired task [7]. In some cases, therapist or 'an agent behind the curtain' controls the robots due to lack of automatic perception of signals and spontaneous response to the emotional cues, consequently making less autonomous human-robot interaction. There is a need of autonomous and data-driven machines that can determine patient behavior and react accordingly [12]. Furthermore, these systems are heavily relying on both audio and video sensors input for making stronger relation. However, robots placement to aid TBI patients in a home or in a specialized neuro-center, face certain additional obstacles that are necessary to be considered. These include the patients' non-cooperative behavior, inappropriate responses and inability to express their emotions. This is due to the nature of

the condition like stroke or accident, resulting in damaged sensory motor control and reasoning skills, along with restricted muscle movements due to paralysis [**?**, 13]. However, these challenges can be different from patient-to-patient, depending on the nature and severity of the injury, producing speech inhibition, partial or complete paralysis, involuntary body movements, abrupt emotional changes, aggression, lack of consciousness or attention and varied emotion elicitation [14]. Therefore, we aim to exploit only visual signals for system generalization for TBI patient's emotional analysis through facial expressions.

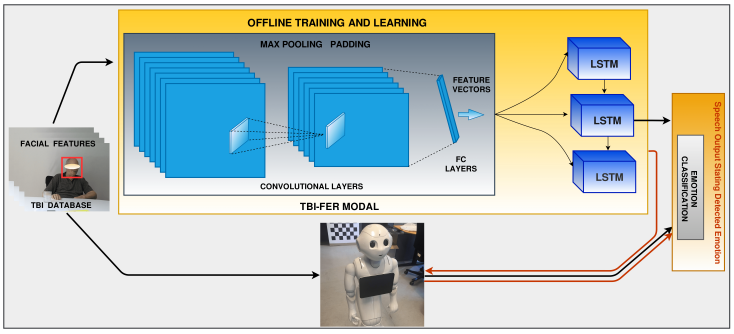


Fig. F.1: Block Diagram of Deep Trained TBI-FER Model and Pepper's Built-in Facial Expression Recognition for Emotion Classification. Black-arrow represents FER through built-in Pepper robot whereas red-arrows represents deployment of TBI-FER modal

Current Facial Expression Recognition (FER) systems are largely based on Convolutional Neural Networks (CNN) for feature extraction and classification as they provide state-of-art results for face recognition [15, 16], facial expression recognition [17–20] and emotional states identification [21, 22]. Their results are highly accurate on healthy people and in controlled conditions. However, this high accuracy is still yet to be achieved with challenging environmental conditions such as large pose variation, low illumination, and on data sets of people with limited expressions like TBI patients. In addition to that, remarkable achievements has witnessed in machines analysis of human emotions, but there are still noticeable challenges that are needed to be addressed in order to involve robots into daily interfaces like social, physical or cognitive activities in real-world scenarios. Some of the major challenges are as follows:

- The wide range of datasets available for FER are collected in laboratory and controlled conditions with little or no pose variations, frontal views, without occlusion, stable illumination and with cooperative subjects. Undoubtedly, such luxuries are not present in real-world applications. Systems trained on such data do not perform well in real-time with real subjects.
- Currently available datasets have FE of healthy people who are mostly cooperative and sometimes produce induced-expressions as compared to TBI patients who have impaired skills, and quite varied and limited expressions due to facial paralysis [**?**, 13]. Additionally, induced FEs that are produced consciously largely alter from natural involuntary emotions.

- Most of existing intelligent systems are trained on databases where expressions are clear with little variance on the vast majority of all 6 basic expressions such as happiness, sadness, fear, angry, surprise and disgust. However, in case of TBI patients classification of 6 basic expressions (7 including neutral) is quite complex as TBI patient's expressions are not easily distinguishable except for one or two and all expressions are not very common. Hence, SARs trained on these databases needed to be customized as these special subjects behave and respond differently than healthy people.

In this research article, we intend to address the aforementioned complexities and limitations in TBI-human-robot interaction by the utilization of a TBI-patient database, which is a collection of multimodal data annotated by TBI-patient's care givers, experts, physiotherapists and doctors. This database is a collection of TBI facial images for spontaneous expression analysis, captured in an entirely unconstrained, real-world environment. It contains the events of natural interactions of subjects of diverse background and age groups in three scenarios of cognitive, social and physical rehabilitation activities. We used this database to develop a deep trained model (TBI-FER Model), composed of Convolutional Neural Network and Long Short Term Memory Networks (CNN-LSTM) to exploit the spatio-temporal information of the TBI subjects. This TBI-FER Model is dedicated for FER of TBI patients that can be integrated with SAR robot, like the Pepper robot for effective human-robot-engagement-research. We performed the classification of 6 basic expressions through the TBI-FER model validated on the TBI patient database as well as the Extended Cohn-Kanade (CK+) (healthy people) database [23]. We also present the hypothesis that our proposed model will outperform the pepper robot built in model. Furthermore, the Pepper robot built-in FER model is employed on both healthy and TBI patient databases and a FE variance analysis is made.

The rest of this article is structured as follows. Section 2 presents the related work including social robots and facial expression recognition. Section 3 describes the methodology for the TBI-FER model training with CNN-LSTM and FE identification through the Pepper robot, and Section 4 describes the experiment and its results. Finally, Section 5 presents the discussion and concludes the paper.

2 Related work

2.1 Social Assistive Robots

In recent years, there has been a growing interest in providing assistance and services to people for physical or cognitive rehabilitation, social interactions and many other health care applications with the help of special robots, categorized as social assistive robots (SAR) [24]. SAR are extensively purposed for monitoring and assisting elderly people in activities of daily life (ADL) in smart homes. Paro, a pet robot resembling a baby seal, has shown positive results for pet therapy by reducing stress in residents in care centers [25]. This also resulted in increased social interaction between residents. Similarly, Roball, a mobile SAR with an IR sensor for touch detection, improved the social interactive behaviour among kids suffering from Autism Spectrum Disorders

(ASDs). Roball has encouraged the kids to play with trainers, therapists, and family members [26]. AIBO (Artificial Intelligence roBOt), an autonomous entertainment robot, was proved effective in enhancing social interaction as well as in aiding mental therapy [27]. AIBO uses touch, audio, vision and thermal sensors to perceive information. A personal assistant robot, Philips iCat, used as a companion, motivator and educator, performed roles of engaging, fostering and instructing [28]. iCat uses vocal emotional expression as well as facial emotional expressions. Another type of robot architecture that integrates the domains of robotics, medicine, psychology, social, cognitive as well as interactive fields is HealthBot [29]. This robot was designed to help the elderly, monitoring their health status and detecting falls. In addition, there is an extensive research on assistive and interactive robotics focusing on the rehabilitation of the elderly and people who suffered stroke [30, 31]. The mentioned companion robots aid in ADL [32] and engaging socially for the purpose of assistance and recovery to improve life quality [29, 33, 34] in the field as well as in lab. Sophia, one of the most advanced humanoid social robot can display expressions similar to humans to build trust and aid humans towards a better life and design smarter homes [35]. She has the ability to process visual, emotional and conversational data. Sophia incorporates Gardner's multiple intelligences [36] into her cognitive architecture. Sophia has also been used as a meditation consultant, giving step by step instructions to help people feel better in lab environment but Sophia has not been placed in field with real subjects. Additionally, these robots utilize different perceived signals such as voice, touch, gestures, signals through IR, RGB, thermal and depth cameras, subject motion tracking, force sensors, and many other indicators to perform their tasks.

Mabu, the intelligent and socially interactive personal health care companion, looks after the patients at home, and mainly reminds them about their medication [37]. Mabu emotionally engages with patients, and evolves its relationship over time by tailoring its conversation by adopting behavior psychology using Artificial Intelligence (AI) algorithms [38]. It also focuses on keeping the patients healthy by constantly monitoring their health and sending encrypted data to a personal doctor if required. Moreover, it actively involves its patients in therapies as prescribed by the doctors. One of the major features of Mabu is active involvement in its speech with patients and the ability to augment its psychological and physiological models to generate new conversational models with the aim of long-term health care [37, 38].

SoftBank robotics have developed NAO [39] and Pepper [40], which are high performance humanoid robots for research and education purposes with the ability to process a wide range of expressions and gesture information. Pepper is equipped with several sensors, but most importantly two 2D and one 3D cameras, which can easily be accessed by its SDKs. Due to its cameras and sensors the Pepper robot can recognize, track and turn while following faces. It also has a preset FER algorithm. The comparison of the discussed robot's input modalities and re-learning capabilities is presented in Table F.1 whereas their illustration is presented in Figure F.2.

Sophia Mabu Pepper NAO iCat HealthBot (top) PARO (bottom) AIBO

Fig. F.2: Famous SAR robots

Table F.1: Robots' Input Modalities And Re-Learning Capability

Robot	Audio Input	Video Input	Tactile Input	Adaptive Re-learning
Sophia [35]	Yes	Yes	No	Yes
Mabu [37]	Yes	Yes	Via tablet	Yes
Pepper [40]	Yes	Yes	Yes	Not by default
NAO [39]	Yes	Yes	Yes	No
iCAT [28]	Yes	Yes	No	No
HealthBot [29]	Yes	No	Via tablet	No
PARO [25]	Yes	Via light sensor	Yes	No
AIBO [27]	Yes	Yes	Yes	No

2.2 Deep Learning Approaches for Facial Expression Recognition

In the aforementioned robots, different sensors have been integrated to achieve efficient human-robot interaction but in our case we would like to rely only on visual information so that the robot can communicate and recognize the emotions of TBI patients effectively, regardless of their speech and locomotion disabilities. It is observed that human emotions are mostly recognized by facial expressions [41, 42]. In order to identify emotions accurately, face and Facial Expression Recognition (FER) approaches have been evolved from holistic, local-feature-based like Gabor or Local Binary Pattern (LBP), learning-based-local descriptors (shallow methods) to deep learning (DL) methods [43]. Traditional methods failed to address certain challenges when researchers moved towards automated and unconstrained FER in challenging conditions. In 1990's, the holistic approaches dominated the FR community with certain low-dimensional representations inferences like linear subspace, sparse representation and manifold approaches [44, 45]. However, these holistic methods failed when exposed to uncontrolled facial changes, different from prior assumptions. This lead to rise of local features based facial recognition methods involving Local Binary pattern (LBP) [46], Gabor [47], SIFT, HOG and other high-level dimensional representations [48]. Unfortunately, these handcrafted features could not address the unique characteristics and denseness of facial features. Following these limitations, researchers introduced the learning-based-local descriptors for better distinctiveness and compactness. This produced FE accuracy of approximately 95% [42] but this is achieved under controlled conditions with frontal views and high resolution images.

Fig. F.3: Pepper robotic administration for Facial Expression Analysis

However, these shallow methods do not handle well non-linear changes in facial appearance. In real time scenarios, shallow methods have improved the accuracy on challenging unconstrained Labeled Faces in the Wild (LFW) dataset [49] to about 95% [50] in 2010. Alex Net won the Image-Net competition [51], through deep learning methods, such as convolution neural networks (CNNs) with a substantial margins. Similarly, in 2014, DeepFace approached close to human performance (97.53%) on LFW dataset benchmark [49], and acquired state-of-arts performance (97.35%) [17]. All of these experimental evaluations are based on subjects without any expression impairments like TBI patients. Ilyas et al in [52] have exercised the CNN-LSTM architecture to exploit the spatio-temporal information for features classification and mood analysis of TBI patients and achieved an accuracy of 87.97% on challenging TBI database. We have employed the same linear combination of CNN-LSTM architecture to train the TBI database and compared with Pepper robot built-in FER model to have FER performance analysis.

3 Methodology

3.1 Database Development and Training

The main aim of this study is to perform facial expression (FE) and mood recognition of TBI patients in order to enhance the social interaction and assist trainers and physiotherapists with the help of robots. First we accumulated a database in three uniform scenarios namely cognitive, physio and social rehabilitation activities, ensuring the reliability of the database as explained in detail in [13, 52]. This database is comprised of 924 videos taken about 11 participants, each being a maximum of 5 second in length, recorded with an Axis RGB and a Logitech RGB camera during multiple sessions at 30fps, resulting in approximately 140,000 captured frames.

For database training, first various pre-processing techniques like face detection, landmarks detection and tracking by Supervised Decent Method (SDM) followed by Face Quality Assessment (FQA), were applied to guarantee high quality images in Face-Log system. In the next step, this high quality image database is passed through a linear architecture of CNN and LSTM, to extract the facial features with the help of CNN from the input faces of TBI patients and then feed to LSTM to exploit the temporal relation on the basis of extracted features in timely manner. For feature extraction we have fine tuned the CNN with off the shelf pre-trained VGG-16CNN model [53]. Features are obtained as `fc7` layer of CNN with VGG-16 model that is feed into LSTM

model to analyze the performance of combined CNN + LSTM deep neural architecture, resulting in TBI patients' FER model (TBI-FER). For performance evaluation the TBI-FER model is validated on the CK+ database. The general schematic of the robotic architecture executed for FER analysis is demonstrated in Figure F.1.

In order to analyze the FER through Pepper robot, the solution required two distinct operations. Firstly, the NaoQi Python SDK is used to retrieve video from the 2D camera of the pepper robot with frame rate of 30FPS and the resolution to 320x240px. Secondly, subject ID file is created and participants were then asked to sit in front of the robot and make different facial expressions related to the 6 emotions. Figure F.3 is illustrating the procedure of emotion elicitation through Pepper robot.

4 Experimental Results

In order to present the results, first we explain the experimental setup. In terms of experiments, we evaluate both FER models namely TBI-FER and Pepper-FER models on TBI and CK+ databases for emotion recognition as seen in Figure F.4.

4.1 Experimental Setup

The robot was set up to perform FER with its built-in detection algorithm in order to later annotate the recorded one minute videos and to serve as a base for comparing the built-in method (Pepper-FER) to our proposed model. This model is also validated on both TBI and CK+ databases. Pepper utilizes that trained model for live classification on the robot. In order to compare with the TBI-FER model, a connection is established with robot similar to video recording and images are retrieved. The images were passed onto the loaded classification model, and the classified emotion was returned as a string. The information was used to be pushed to the robot through another initialized service converting text to speech (TTS). As a result, the robot was capable of reporting the participants' emotions through TTS with our proposed FER model.

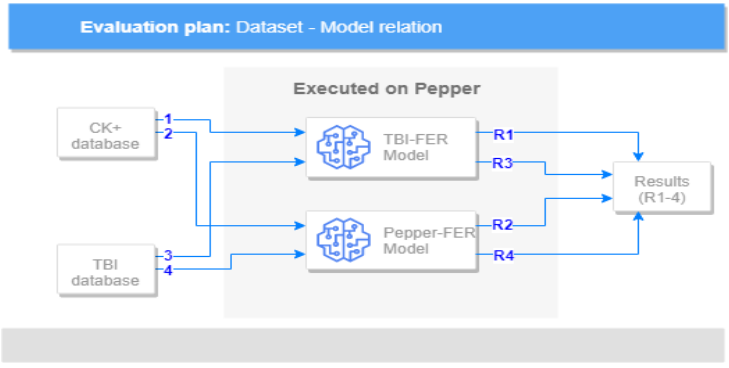


Fig. F.4: Evaluation of FER Models on TBI and CK+ Database

Table F.2: Confusion Matrix of 6 Basic Expressions through TBI-FER model on TBI Patient's Database

	Neutral	**Happy**	**Angry**	**Sad**	**Fatigued**	**Surprised**
Neutral	**88**	3	2	14	2	1
Happy	4	**82**	2	3	2	7
Angry	2	2	**85**	5	6	1
Sad	12	1	4	**78**	11	1
Fatigued	7	1	5	2	**67**	9
Surprised	2	21	3	2	6	**71**

4.2 TBI-FER Model Analysis

In this section, we discuss the training of our system on the TBI patient database and its validation of the results for 6 basic expressions. It is evident that the neutral expression has the highest, 88% accuracy, as shown in Table F.2. This is due to the fact that neutral is the most common expression in TBI database. Although, in most cases TBI neutral expression is most likely recognized as sad for healthy people. Fatigued or stress expression exhibits the lowest accuracy in the validation of this FER model. This is due to the unbalanced data set, which is a result of the difficulty of acquiring this type of data because of stressed or non-cooperative participants. On the other hand, when this TBI-FER model is employed on the CK+ database for identification of expressions, it is shown that the CK+ database results are much better compared to the TBI patient one due to the reason that latter database is mainly of high quality images with frontal faces. Comparatively, in case of the TBI patients, there is challenge of working with non-frontal faces. FE of neutral, angry, sad, happy, surprise and fatigue are identified accurately up to 91%, 88%, 87%, 85%, 84% and 82% respectively as illustrated in Table F.5.

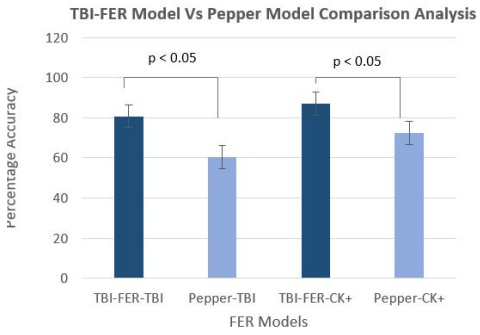


Fig. F.5: Comparison of FER Modals on TBI Patients and Healthy Subjects

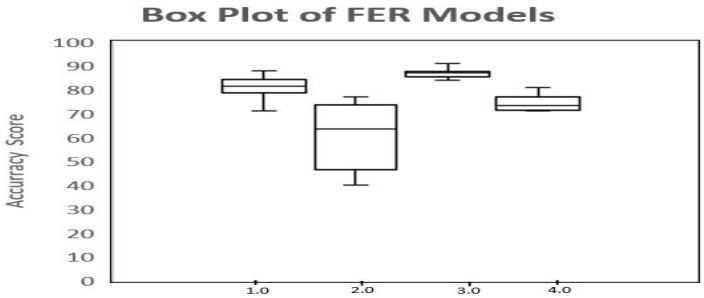


Fig. F.6: Box plot of the FER accuracy score. On the x-axis 1.0 is TBI-FER model on the TBI database, 2.0 is Pepper-FER model on the TBI database, 3.0 is TBI-FER model on the CK+ database, 4.0 is Pepper-FER model on the CK+ database

4.3 Pepper built-in FER Model Analysis

For the classification of emotions through Pepper, its built-in FER model is implemented on both TBI patient and healthy people database. It is observed that Pepper identified the surprise emotion from TBI patients with an accuracy of 42% as opposed to 71% for CK+ database as demonstrated in Table F.3 and F.4 respectively. This can be due to the varied and limited surprise elicitation from TBI patients due to stroke impact. Furthermore, Pepper identifies neutral expressions of TBI patients with only 42% accuracy with sad and neutral expression overlapping, proving that TBI patients' neutral expressions are more likely recognized as sad ones. Experts have annotated the patients' expressions as neutral since their ability to display emotional signals is disturbed due to injury, and during post stroke rehabilitation they exhibit depression and negative emotions more often than positive ones [13, 52]. It is also observed that the Pepper robot failed to identify fatigue expressions due to technical limitations.

In order to determine which FER model is significantly more accurate, we have conducted a student's t-test on the TBI-FER model and the Pepper built-in model, where variance is approximated for each of the model. For t-test each of the model has to follow the normal distribution and this validated by Q-Q plots and K-S normality tests. We conducted t-tests on two separate databases for each of the FER model. As seen in Figure F.5, for TBI database, the t-value comes out 2.54 with a p-value 0.023. Thus, the null hypothesis can be rejected and we can conclude that the TBI-FER model is significantly more accurate. By studying the box plot in Figure F.6, it can be seen that TBI-FER score is greater than Pepper-FER score, it can be concluded that TBI-FER model has higher accuracy than the Pepper-FER model on the TBI-database. Similarly, when examining the CK+ database for FER models accuracy, the t-value comes out 3.17 with the p-value 0.003. Thus, we can conclude that the TBI-FER model is also significantly more accurate than the Pepper FER model for healthy subjects. It is also clearly evident in the box plot in Figure F.6, the TBI-FER model has higher score than Pepper-FER model on CK+ database.

Table F.3: Confusion Matrix of Facial Expression Recognition through Pepper-Robot Built-in Model on TBI Patients

	Neutral	**Happy**	**Angry**	**Sad**	**Fatigued**	**Surprised**
Neutral	**42**	0	12	18	x	1
Happy	1	**67**	2	0	x	5
Angry	12	5	**73**	9	x	2
Sad	17	1	12	**76**	x	2
Fatigued*	x	x	x	x	x	0
Surprised	2	2	3	2	x	**42**

* The Pepper robot lacks ability to identify fatigue expressions.

Table F.4: Confusion Matrix of Facial Expression Recognition through Pepper-Robot Built-in Model on Healthy People

	Neutral	**Happy**	**Angry**	**Sad**	**Fatigued**	**Surprised**
Neutral	**59**	1	5	7	x	1
Happy	14	**74**	2	2	x	23
Angry	11	3	**78**	4	x	2
Sad	17	1	9	**81**	x	3
Fatigued*	x	x	x	x	x	0
Surprised	2	15	7	2	x	**71**

* The Pepper robot lacks ability to detect fatigue expressions.

5 Conclusion and Discussion

In the general context of FER and social interaction of TBI patients, this paper has presented a robotic framework to identify the FE and emotional signals of TBI patients specifically by introduction of customized deep trained model to meet the requirements of a specialized scenario. To do so, two FER-models, customized TBI-FER model and Pepper-FER model are compared, and their performance is analyzed. For this purpose, TBI patients database was collected in three uniform scenarios, than deep trained model composed of linear combination of CNNs and LSTM is devel-

Table F.5: Confusion Matrix of 6 Basic Expressions through TBI-FER Model on CK+ Database

	Neutral	**Happy**	**Angry**	**Sad**	**Fatigued**	**Surprised**
Neutral	**91**	2	3	5	1	1
Happy	3	**85**	2	3	2	4
Angry	2	2	**88**	5	6	2
Sad	5	1	4	**87**	12	2
Fatigued	5	1	5	3	**82**	2
Surprised	5	4	1	2	6	**84**

oped to identify the FE and mood of TBI patients. This model is compared with the Pepper robot built-in FER model and FER accuracy is determined using objective assessment methods. Objective evaluation method is used by analyzing facial expressions on test subjects. The results demonstrated that TBI-FER model has significantly higher performance as compared to the Pepper-FER model, on both TBI database and CK+ database (healthy subjects). Furthermore, individual expressions are more pronounced by TBI-FER model, this cross validates the previous results. So in order to place the Pepper robot with TBI patients, it is essential to use customized trained model for more meaningful interaction. Facial expression recognition has proved to be a vital tool to evaluate the mood of subjects in non-obtrusive manner for enhancing social interaction. Therefore, the Pepper robot can use these self-trained models, in our case a TBI-FER model. This can lead to behavioral adaptation of the robot in accordance with patient mood, similar to the implementations of Mabu and Sophia [35, 37] but with less cost and computational power.

References

[1] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2016, pp. 3723–3726.

[2] A. E. Eiben, "Grand challenges for evolutionary robotics," *Frontiers in Robotics and AI*, vol. 1, p. 4, 2014.

[3] D. Feil-Seifer and M. J. Matarić, "Socially assistive robotics," *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 24–31, 2011.

[4] M. J. Matarić, "Socially assistive robotics: Human augmentation versus automation," *Science Robotics*, vol. 2, no. 4, p. eaam5410, 2017.

[5] B. Davies, "Robotic surgery–a personal view of the past, present and future," *International Journal of Advanced Robotic Systems*, vol. 12, no. 5, p. 54, 2015.

[6] S. P. DiMaio and S. E. Salcudean, "Needle steering and motion planning in soft tissues," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 6, pp. 965–974, 2005.

[7] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, p. eaao6760, 2018.

[8] P. Chevalier, J.-C. Martin, B. Isableu, C. Bazile, and A. Tapus, "Impact of sensory preferences of individuals with autism on the recognition of emotions expressed by two robots, an avatar, and a human," *Autonomous Robots*, vol. 41, no. 3, pp. 613–635, 2017.

[9] P. Chevalier, "Social personalized human-machine interaction for people with autism," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 229–230.

[10] J. Abbasi, "In-home robots improve social skills in children with autism," *Jama*, vol. 320, no. 14, pp. 1425–1425, 2018.

[11] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *arXiv preprint arXiv:1901.02839*, 2019.

[12] S. Harker, "Applied behavior analysis (aba)," *Encyclopedia of Child Behavior and Development*, pp. 135–138, 2011.

[13] C. M. A. Ilyas, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," *25th IEEE International Conference on Image Processing (ICIP)*, 2018.

[14] D. T. Stuss and B. Levine, "Adult clinical neuropsychology: lessons from studies of the frontal lobes," *Annual review of psychology*, vol. 53, no. 1, pp. 401–433, 2002.

[15] H. Li and G. Hua, "Hierarchical-pep model for real-world face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4055–4064.

[16] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4362–4371.

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[18] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[19] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[20] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*. Cham: Springer International Publishing, 2017, pp. 151–162.

[21] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49–61, 2018.

[22] J. Wan, S. Escalera, G. Anbarjafari, H. Jair Escalante, X. Baró, I. Guyon, M. Madadi, J. Allik, J. Gorbova, C. Lin *et al.*, "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3189–3197.

[23] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

[24] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *9th International Conference on Rehabilitation Robotics*, 2005, pp. 465–468.

[25] K. Wada, T. Shibata, T. Saito, K. Sakamoto, and K. Tanie, "Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 2005, pp. 2785–2790.

[26] T. Salter, F. Michaud, D. Létourneau, D. C. Lee, and I. P. Werry, "Using proprioceptive sensors for categorizing human-robot interactions," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2007, pp. 105–112.

[27] M. Fujita, "On activating human communications with pet-type robot aibo," *Proceedings of the IEEE*, vol. 92, no. 11, pp. 1804–1813, 2004.

[28] J. M. Kessens, M. A. Neerincx, R. Looije, M. Kroes, and G. Bloothooft, "Facial and vocal emotion expression of a personal computer assistant to engage, educate and motivate children," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[29] C. Jayawardena, I. H. Kuo, E. Broadbent, and B. A. MacDonald, "Socially assistive robot healthbot: Design, implementation, and field trials," *IEEE Systems Journal*, vol. 10, no. 3, pp. 1056–1067, 2016.

[30] S. Ueki, H. Kawasaki, S. Ito, Y. Nishimoto, M. Abe, T. Aoki, Y. Ishigure, T. Ojika, and T. Mouri, "Development of a hand-assist robot with multi-degrees-of-freedom for rehabilitation therapy," *IEEE/ASME Transactions on Mechatronics*, vol. 17, no. 1, pp. 136–146, 2012.

[31] T. Yokoo, M. Yamada, S. Sakaino, S. Abe, and T. Tsuji, "Development of a physical therapy robot for rehabilitation databases," in *2012 12th IEEE International Workshop on Advanced Motion Control (AMC)*, 2012, pp. 1–6.

[32] J. Saunders, D. S. Syrdal, K. L. Koay, N. Burke, and K. Dautenhahn, "x201c;teach me x2013;show me x201d; x2014;end-user personalization of a smart home and companion robot," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 27–40, 2016.

[33] K. Swift-Spong, E. Short, E. Wade, and M. J. Matarić, "Effects of comparative feedback from a socially assistive robot on self-efficacy in post-stroke rehabilitation," in *IEEE International Conference on Rehabilitation Robotics (ICORR)*, 2015, pp. 764–769.

[34] J. Fan, D. Bian, Z. Zheng, L. Beuscher, P. A. Newhouse, L. C. Mion, and N. Sarkar, "A robotic coach architecture for elder care (rocare) based on multi-user engagement models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 8, pp. 1153–1163, 2017.

[35] B. Goertzel, J. Mossbridge, E. Monroe, D. Hanson, and G. Yu, "Humanoid robots as agents of human consciousness expansion," *arXiv preprint arXiv:1709.07791*, 2017.

[36] L. Helding, "Howard gardner's theory of multiple intelligences," *Journal of Singing*, vol. 66, no. 2, p. 193, 2009.

[37] M. J. Johnson, M. A. Johnson, J. S. Sefcik, P. Z. Cacchione, C. Mucchiani, T. Lau, and M. Yim, "Task and design requirements for an affordable mobile service robot for elder care in an all-inclusive care for elders assisted-living setting," *International Journal of Social Robotics*, pp. 1–20, 2017.

[38] C. Datta, "Programming behaviour of personal service robots with application to healthcare," Ph.D. dissertation, ResearchSpace@ Auckland, 2014.

[39] A. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: pepper: the first machine of its kind," *IEEE Robotics & Automation Magazine*, no. 99, pp. 1–1, 2018.

[40] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi, "Pepper learns together with children: Development of an educational application," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 270–275.

[41] Y. Wu, H. Liu, and H. Zha, "Modeling facial expression space for recognition," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2005, pp. 1968–1973.

[42] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, second edition ed. Springer London Dordrecht Heidelberg New York: Springer, 2011.

[43] M. Wang and W. Deng, "Deep face recognition: A survey," *arXiv preprint arXiv:1804.06655*, 2018.

[44] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 711–720, 1997.

[45] W. Deng, J. Hu, J. Guo, H. Zhang, and C. Zhang, "Comments on" globally maximizing, locally minimizing: Unsupervised discriminant projection with application to face and palm biometrics"," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1503–1504, 2008.

[46] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2037–2041, 2006.

[47] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467–476, 2002.

[48] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 786–791.

[49] L. J. Karam and T. Zhu, "Quality labeled faces in the wild (qlfw): a database for studying face recognition in real-world environments," in *Human Vision and Electronic Imaging XX*, vol. 9394. International Society for Optics and Photonics, 2015, p. 93940B.

[50] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3025–3032.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[52] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[53] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

Part IV

Dissemination Activities

Paper G

Developing a user-centred Communication Pad for Cognitive and Physical Impaired People

Chaudhary Muhammad Aqdus Ilyas, Kasper Rodil and Matthias Rehm

The paper has been published in the
Proceedings of 8th EAI International Conference: ArtsIT, Interactivity Game Creation, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, (LNICST, volume 328),
doi:10.1007/978-3-030-53294-9$_9$
2019.

The layout has been revised.

Abstract

It is always challenging for people with disabilities, particularly having speech inhibition to communicate. In this research article, we explored the case study of the resident at the neurological centre, having a complication in conveying messages due to physical and speech paralysis. For making effective communication, we have developed a user-centred communication pad where the resident needs to swipe a finger on the pad with printed alphabets and digits (we called it communication pad). A camera placed over the communication pad detects the finger movement of the resident and extract the message to display on the computer screen or the tablet. Our tracking method is robust and can track the fingers even in varying illumination conditions. This paper also covers the main steps of design methods with various design prototypes and its user feedback. Result analysis of different design modules and user experience evaluation shows that our designed system has provided independence and convenience to the resident in conveying a message successfully.

1 Project Introduction

Since 2015 we have been working with a national neurological centre (hereafter neuro centre) with a focus on co-designing various technical systems enhancing capability for the individual residents. As these residents are unable to recover from their life-altering impairments fully, the centre provides full-time care to them and aid in organizing and supporting activities of daily living (ADL). The project collaboration aim is to investigate where technological innovation can assist residents and staff members with fulfillment of rehabilitation activities - including enhancing individual self-control and improvement of quality of life [1–3].

One of the overall design (and research) challenges is the unique (and highly diverse) nature of the cognitive abilities of residents (for instance, apraxia and aphasia). Due to the severe and diverse conditions, the residents require assistance even for small chores. To list a couple of examples, some of the residents are fully paralyzed and bound to wheelchairs or beds, some residents have lost all speaking ability, and some have minimal short-term memory or attention spans (in some cases less than two minutes). All residents embody a combination of these impairments, but the common characteristic is that they all became impaired late in life. These conditions reflect a significant alteration of the functionality of the individual - in many cases leading to depression and general loss of life quality perceptible as a decrease of "self-control, self-worth, privacy and independence" [3].

While the primary task of the neuro centre is to provide round the clock caregiving, it also encourages technical solutions addressing the needs of these residents for specific task assistance as these residents are heavily relying on staff support even for personal and private matters. It is important to stress that it is not only a budgetary manoeuvre, but there is a grounded wish for the residents to have as much self-control as possible. Thus a major research strand orbits; how to enable designing for diversity with an inclusive design approach - such as Participatory Design.

Some companies who are working with the neuro centre furnish technical support related to rehabilitation activities but with little to no consideration of personal

challenges and abilities. Most of their products are designed for rehabilitation purposes only with highly generic solutions, thus making them of little to no use for these residents. Each resident has a unique and individual challenge, for instance, one resident has a problem with remembering, he frequently forgets forcing staff to remind him time and time again about even very basic tasks. This cycle of reminding and forgetting often leads to frustration on both parts.

We have been part of a variety of different projects at the neuro centre over the years and after a series of consultation meetings (demonstration of prototypes, group talks, socialization) with staff members and residents, there was consensus to focus a project on making customized, and human-centred functional social robots to enhance independence and quality of life. The project demonstrates a well-meant objective of empowering the residents to respond to their everyday challenges and give a voice to those who are neglected or technically limited to be part of otherwise off-access traditional system development.

Thus the inclusion into design is cardinal and a priority that the residents provide input during design sessions and contribute to the aesthetic and functional properties of the systems. This deliberate inclusion has so far provided the residents with a visible sense of ownership. As an example, in some situations, the residents suggested making the design to closely reflect the portrayal of their favourite movie character or other more personal traits. What was initially the project, became an umbrella for several individual projects. Albeit being very different in function and aesthetic, they all followed the same development model rooted in problem-oriented development. The first phase was best characterized as an ethnographic approach into the life world of the resident and the particularities of their situations requesting a technical solution. Following this phase resembled a typical collaborative sketching/illustration on paper phase whereby ideas were externalized (for instance by using cardboard). The last phases involved prototyping with 3D printers, assembly using electronics and always with several sessions together with the resident. These social robots were from the beginning customized for and with a specific user - one system for one resident.

1.1 Case Study

People with motor, speech and hearing inhibitions face severe difficulties in conveying their messages traditionally (for instance using sign language). In many cases, they are dependent on Augmentative and Alternative Communication (AAC) technologies so there is always a need of a specific communication system, for instance, one that can track hand or finger gestures. Thus, such a camera vision system able to transcribe finger or hand movement or sign language into text or speech would, conceptually, be useful for productive interaction (reliable and fast).

In this case study, the resident is suffering from speech inhibition and paralysis and is used to communicate with staff through an analog communication tool, a big-sized letter-board, with digits and numbers printed as illustrated in the Fig G.2. First of all, the board with printed numbers is quite big, making it unfit to use it in all situations. For instance, if a resident require assistance while travelling or even social communication outside the resident's apartment, he is not able to use this tool as it is often only available in his apartment. The actual one-to-one communication requires

staff members to point on various letters to overtime construct words and sentences and vice versa. Pointing out the letters on the board is very tedious for the resident as well as for the staff member, and most of the time, it leads to confusion. Therefore, residents and staff members have to repeat the process many times over to exchange even basic information.

In addition, this process is exposing the privacy of the resident to the staff members. The resident is like the other residents staying at the neuro centre permanently and can not communicate freely with visiting friends or family members without the presence of staff. Most visible is the problem when the resident exchanges text messages with family and friends. The staff member will have to (besides decoding the intended message on the board) type the message on the resident's mobile phone and afterwards return to read it out loud. In some cases and because of this troublesome process, the resident is hesitant to communicate with ex-situ family members.
We decided to address the challenges both at the vector of the physical system design side and at the vector relating to the convenience and privacy issues for the resident. Having these factors in mind, we devised a proof of concept vision-based system called "visual communication pad" (Vis-Com pad). The Vis-Com pad concept was scoped around making a vision-based real-time text recognition system that automatically detects the finger movement over the pillow-board (letter-board) to infer the text message that can be displayed to a screen or sent to the receiver through a communication device; such as a mobile phone. At the end the Vis-Com pad has enabled the resident to convey a message without the intervention of staff. Section 3 provides the details of system designing and implementation. Before addressing the design and implementation of the system, we will address the technical landscape on which the system rests.

2 System related literature

Most of the camera-based systems, which use a hand as the basis for non-verbal communication, conform to a sequence of steps: detection and segmentation; tracking and feature extraction; and finally classification. The first step is detection and segmentation of hand or fingers in the field of view (FoV) of the camera. In the next step, the detected hand is tracked, and visual features are extracted. In the last step, spatiotemporal data that is extracted in the previous step are grouped and assigned specific labels.
The primary aim of vision-based hand gesture recognition systems is the clarification of the semantics of hand movement, posture attribution or bodily expression cues [4]. These signals play an integral part in the understanding of the message. It is also necessary to process this information in real-time and to enable the system to respond accordingly. We can distinguish the gesture recognition systems based upon the input data type such as RGB, thermal or depth; methods used to process the input information (employment of various segmentation, feature representation and classification approaches) like geometric, graphical or machine learning or deep learning approaches; and application of system with static or kinetic background [4–6].

In order to identify hand gestures and finger movement, various sensors can be employed like Microsoft Kinect camera, IR sensor or RGB sensor. Modern technolo-

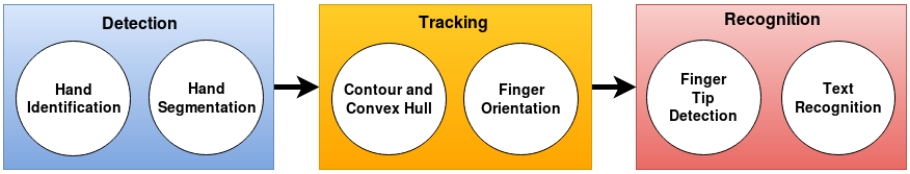


Fig. G.1: System Flow Process for Visual Communication Pad

gies for hand gestures are incorporated with information of depth and distances captured by a 3D camera. The Microsoft Kinect camera can provide depth information at low cost and is a central part of hand gesture recognition systems. Raheja et al. [7] used Kinect camera for gesture recognition in a contact-less manner and tracked fingertips and demonstrated 99% accuracy with extended fingers. Depth based systems have achieved the accuracy of 99.07% whereas RGB-based systems are accurate up to 99.54% and combined modalities have demonstrated the 99.54% of accuracy [8]. This suggests that RGB based systems are good enough for hand gesture recognition systems as there is not a significant difference in accuracy between RGB and depth systems.

Vision based systems make use of various body features for hand and finger recognition. Some researchers have applied graphical models for visual object recognition and tracking, graphical models with depth information and exploiting the bag of 3D points method [6] [9] [10]. Some researchers have exploited skin texture and color information to detect hand or fingers [11], hand shape [12] [13], pixel values [14], 3D hand models [14] [12] [15], and utilization of hand motion knowledge through boosted histograms [16]. Muhammad et al. proposed a hand gesture system which detects the hand and identifies its center and thus the hand movement is tracked with the position of the hand [17].

Each technique has its embedded advantages and disadvantages and selecting the most appropriate one can not be done without contextual understanding. As we will demonstrate, not all decisions are guided by technical performance but is instead a combination of various factors. After all, the system is not intended as a pure technical construction for a lab experiment, but intended to function in a 'wild' setting intertwined with both social relationships, contextual factors (such as lighting conditions) and individual technical abilities.

3 Implementation of the Vis-Com pad system

The formulation of the Vis-Com pad system was informed as a combination of technical possibilities, and from the field informed contextual factors and human factors (such as lighting and the complex set of abilities of the resident).

Modern vision systems are incorporated with RGB and depth sensors, but we chose only RGB sensors due to the following reasons. First, there is not too much difference in the accuracy of two sensors for hand gesture and finger identification, as mentioned by [8]. Secondly, the use of the Microsoft Kinect camera was imposing

Alphabets on White board Pillow Board

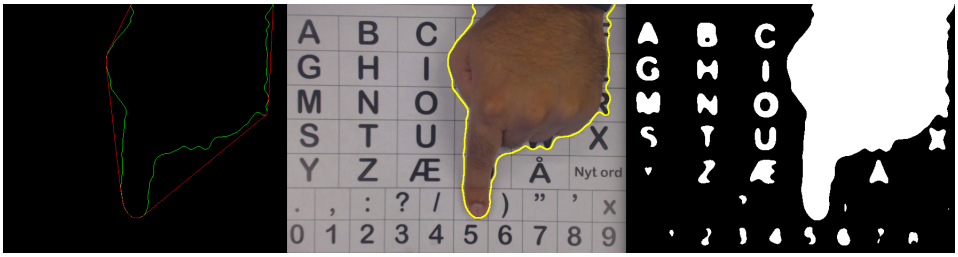
Fig. G.2: Design progress from white board to Pillow board

bulkiness to the system. In short, we chose the RGB sensor by keeping the device employment precision, reliability, weight and size, and suitability for the resident use. In terms of software development, We have employed geometrical descriptors to segment out hand features as the hand is the closest object to the camera. Threshold and region-growing techniques are used to identify hand features as in [11] [13]. In the next phase, we applied the contour, and convex hull techniques to detect the shape and boundary of hand and fingers as illustrated in the Fig G.3. We also applied the thinning algorithm to detect the fingertips. We did not apply hand silhouettes as shape descriptors as it is erroneous when fingers are folded [13] [12]. Our method is relatively close to [17] with fixed coordinate values of the letter board, where finger movement is tracked. For construction of sentences, the finger position over the communication pad is identified, tracked, and labels are assigned based on the spatiotemporal data.

The first step towards the development of Vis-Com Pad is the reduction of the big-sized board to 42-by-30 cm board fixed on top of a pillow as seen in the figure G.2. This "pillow-board" is used to train the resident to move fingers over different letters/numbers, and staff members infer the message and write on a whiteboard or speak verbally to confirm intended meaning. This process helped in two ways; firstly, it involved some physical movement of the hand, considered as physiotherapy for the disabled resident at a basic level. Secondly, it provided the resident with added freedom and motoric ease.

We decided to automate this finger tracking and text recognition process by the installation of the camera at the top of pillow-pad despite the proximity of the camera and letter board. This camera installation caused additional computer vision challenges such as illumination issues, false detection, and occlusion problems that are discussed in detail in section4. Additionally, subjects with paralysis may have issues with placing or pointing fingers at one alphabet/digit at a time. On the other hand, installation of the camera with pillow-pad-arm created issues of inconvenient use due to size and weight of pillow-pad and pillow-arm. While designing the pillow-pad, we considered the size of alphabets or digits should be big enough so that the staff member can see it from a far distance. It was designed for the resident training through a staff member. However, in the final prototype, it was not required when text recognition is carried out by the camera. After careful observation and user's input, we decided to make an A4-size letter board with only 29 letters and 0-9 digits printed on one side of it and the other side with an additional few emojis.

We also decided to install a ring of light around the camera to counter the light-



Background substraction Hand detection on Letter board Hand segmentation

Fig. G.3: Hand Detection and Segmentation process: Central figure illustrates the identification of hand over the letter board; left figure illustrates the background substraction along with contour (green line) and covexhull (red line) application; Left figure is demonstrates the application of thresholding on detected hand

ening issues. This light-ring can change its intensity if required, or the resident wants to communicate at any time relying less on the room light. Furthermore, in the final phase, this letter board is printed over plastic due to lightweight and preventing it from potential damages due to exposed use. To make this portable, we used a Raspberry Pi connected with a camera for tracking finger movements. The tracked information is sent wirelessly to screen or monitor to display the text. This system provides the facility to edit word or sentence before finalizing it or sending it to the intended user to ensure preciseness of the text. Technical details of finger tracking and text recognition system are presented below.

1. **Hand Detection and Segmentation** The first step is the detection of hand and its separation from the background, in our case, it is a letter board. As our background is static, background subtraction is applied to segment out a hand. We applied thresholding to segment the hand from the background, assigning a particular threshold value to the hand region, as illustrated in Fig G.3.

   For the segmentation, we assume that the subject uses only one hand at a time, and it occupies a significant portion in the Field of View (FoV) of the camera. Furthermore, the hand is closer to the camera, and there is no occlusion between camera and letter board besides the hand. There is a small distance between the communication pad and its camera-arm that is approximately 36cm.

2. **Hand Tracking** At this stage, hand motion is tracked over the letter board. Contour and convex hull techniques are applied to draw contour lines around the hand blob and then convex hull around the contour of the hand like an envelope. When a subject moves his hand over the letter board, corresponding segmented hand regions are identified in previous and current frames.

3. **Hand Feature Extraction and Finger Identification** The position and orientation of the hand are determined after the identification of hand regions. As the letter board has printed letters and numbers with a specific orientation, thus hand orientation should be parallel to letter board orientation. However, dealing with paralyzed persons, it is difficult for them to keep their hands in an

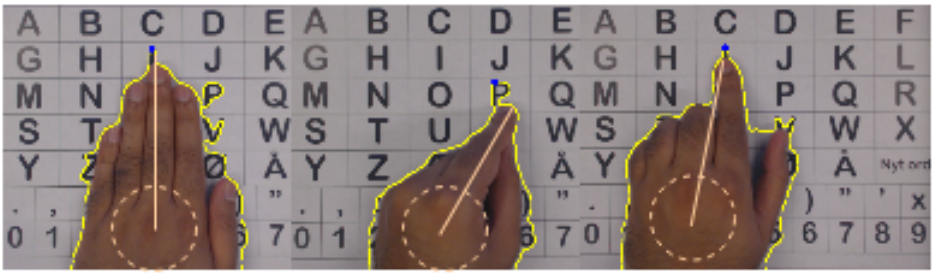


Fig. G.4: Identification of a pointing finger by measuring the maximum distance from center of the palm to the fingers

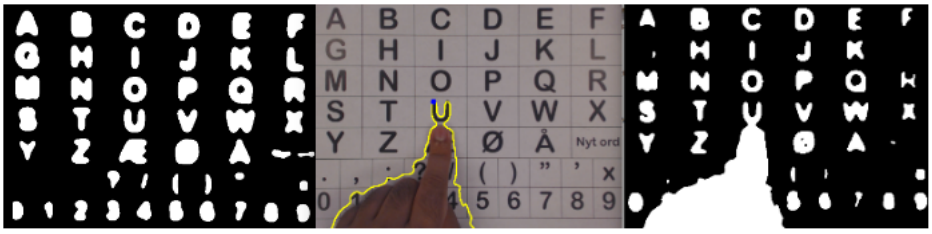


Fig. G.5: Letter recognition process with letter board thresholding and fingertip allocation over letter board to identify the letter.

upright position. The hand position was determined by three directions; up, left and right. The finger is then identified and tracked, and we used fingertip of the subject as the input pointer just like a mouse pointer. For the precise allocation of the finger over letter board, counter tracing algorithm is utilized that detects all the fingertips. The pointing finger and fingertip are determined in two steps. In the first step, the centre of the palm is identified with finger directions. In the next step, the maximum distance from the palm-centre to the fingertip is calculated to identify the pointing finger, as illustrated in figure G.4.

4. **Text Recognition** Text recognition is done in two parts. In the first phase, the letter board is processed with thresholding to identify individual letters and numbers. As their positions are fixed, so their coordinates are stored. In the second step, the fingertip of the pointing finger is located over the letter board coordinates to identify the text.

In the neuro centre, we tried to implement this Vis-Com pad system by utilization of the mentioned computer vision techniques and modified the design parameters. As this project is implemented in a real setting, many challenges have been faced, and various prototypes were tested and designed iteratively.

4 Loops of evaluating the Vis-Com pad

In order to evaluate the system, the basis consisted of three questions all typical resonating conversation. In each prototype evaluation, these three questions were asked in Danish "Hvad hedder du (What is your name)?" "Hvor gammel er du (How old are you)?" "Hvad kan du lide at spille (What do you like to play)?" Previously, staff members used to point out letters on the board to construct a sentence and then sought confirmation by the resident, who would nod in agreement or disagreement. The staff members knew the resident name, age and sports-liking so they can quickly infer and write it down for the resident. However, in other real-life scenarios, this approach, as already mentioned, is time-consuming and prone to errors. Therefore, in Vis-com system, instead of a staff member, the camera tracked the hand and finger positions of the resident and registered the alphabets or letter to formulate the sentence for the intended message. We recorded the video of the whole process accounting for the accuracy of letter registration, time of completion, and the number of repetitions to execute the task. Details of each prototype development and evaluation outcomes are presented in the following section.

4.1 Prototype-I: Short description and findings

In the prototype, Vis-Com pad has a wooden arm with 36cm in height and an adjustable camera holder. This camera holder allows the camera to stay at the center of the letter board that is made up of cardboard with a printed sheet of letters and numbers on it. Vis-Com Pad can be placed on the top of the pillow and can be used by the resident in sitting and lying positions. The whole setup was small and portable. The camera is connected to a Raspberry Pi that is fixed at the base. Text can be displayed to the monitor or tablet screen through wireless communication. When we conducted the evaluation, we encounter the following challenges that lead to the development and implementation of the second prototype.

- The Vis-Com Pad is very sensitive to illumination conditions, so with natural light and room light results have variations and miss detection.
- Lighting positions cause the shadow on letter board, which in turn lead to the false convex hull. It is observed that this problem can be avoided if a focused light is installed over the letter pad.
- Resident hand orientation is different than healthy people hand. Therefore, the fingertip location has erroneous results.
- The letter board and the camera arm produces reflections, one contributing reason in false letter detection in the text recognition process.
- Wooden arm with camera holder was a bit bulky, creating some imbalance when placed on the pillow.

4.2 Prototype-II: Short description and findings

We overcame these identified challenges by the introduction of following changes in the physical design and computer vision techniques of the second prototype.

Illumination Issue Hand orientation Issue Shadow Issue

Fig. G.6: Prototype challenges at various stages of testing procedures due to illumination, hand orientation and faulty convex hull formations

- We introduced a light ring made of LEDs to avoid illumination variation like [18], who introduced the external light source while collecting data for hand gestures. In our system, the camera sensor is surrounded by a light ring so that light falls equally on all parts of the letter board. This light-ring installation minimized the false detection and faulty convex hull formation.
- To reduce reflection from the letter pad, we painted the letter pad and camera arm with black color. The letter board remains in white with black printing. This lead to the additional problem of thresholding as letter pad and letters are now of the same color.
- We decreased the length of camera-arm from 36cm to 30cm to reduce the field of view of the camera so that it captures only coordinates of letters and numbers instead of borders. This reduction in size solved the problem of thresholding.
- Due to the unique hand orientation of the resident, we introduced the new method to locate the pointing finger by measuring the maximum distance from the centre of the palm to the direction of the fingers as illustrated in the Fig G.4. In addition to that, we introduce the determination of hand orientation from three sides, namely, left, right, and bottom. This 3-sides checking ensures the right direction of pointing fingers.
- The letter pad and camera arm was bulky. Therefore, it is suggested to change the wooden arm with a light-weight aluminum rod.
- The letter board is made up of cardboard and is not durable. When the resident moves his finger over the letter board, it bends. Thus produces a change in coordinates of letters, resulting in false text recognition.

4.3 Prototype-III: Short description and findings

In the final version of the Vis-Com pad system, we made following design and technical improvements. This system addresses the challenges raised in previous testing procedures.

- Camera arm is replaced with the black-painted aluminum rod to reduce the weight issues of the system.
- Camera arm length is further reduced to 29cm with a fixed camera position, so that camera field of view (FoV) remains inside the border of letter board coordinate system as illustrated in the figure G.7.

Visual-Communication Pad Raspberry Pi LED light ring with controller

Fig. G.7: Visual-Communication Pad with Raspberry Pi, LED ring and intensity controller

- The light intensity of the LED ring is made adjustable through an RF wireless controller. This light intensity controller provides the resident to use Vis-Cam system without depending upon room light during night time. LED ring controller works on day and night mode only to avoid any false text recognition due to illumination variation.
- The letter board of Vis-Com pad is made with dense and light-weight plastic fiber to avoid the bending problem. Plastic fiberboard is more durable and elastic resistant as compared to cardboard.

We analyzed Vis-Com system performance in terms of accuracy of writing script, time and convenience with the analog communication tool at each prototype development as illustrated in Table G.1. In the first iteration or prototype-I evaluation, due to illumination, design and practical implementation issues, the camera did not extract any useful text information. In this test, the staff member inferred the information from the finger movement over the letter board. In the second prototype testing, some of the letters are printed correctly, but could not construct meaningful sentences. In this stage, staff member intervention helped in retrieving the information from the resident. In the third iteration, after addressing the illumination, speed and design issues, Vis-Com system accurately tracked the finger movement over the letter board and constructed the sentences. The resident was able to write the intended sentence precisely with a display on the screen.

The analog system complexity increases with increase in the length of sentence words or characters, due to repetition and re-writing both by the resident and the staff members. A sentence with five-words or twenty-eight characters consumed 59 seconds and 47 seconds with analog and Vis-com system respectively, as demonstrated in Table G.1. The analog system is proved slow and more tedious as compared to Vis-Com system. Also, the resident valued the Vis-Com system a more convenient and efficient tool to communicate. We have analyzed the ease-of-use of the system with scale 1- to -10, with score ten at the most convenient and scored 1 with the most challenging level. Resident rated our system a more user-friendly with a rate of 8.0 as compared to the conventional approach (where the user needs to iterate multiple

Table G.1: Vis-Com system performance evaluation with the analog communication system in the neuro centre

Qustions		Q1 Hvad hedder du (What's your name?)	Q2 Hvor gammel er du (How old are you?)	Q3 Hvad kan du lide at spille (What do you like to play?)
Answers		**Jeg er John (I am Jhon)**	**Jeg er 30 år gammel (I am 30 years old)**	**Jeg vil gerne spille fodbold (I would like to play football)**
No of Words		3	5	5
No of Characters		11	19	28
Writing Time (Seconds)	**Analog System**	32	48	59
	Vis-Com System	14	31	47
Convenience Scale	**Analog System**	6	5	4
	Vis-Com System	8.5	8.5	8.0
Writing Accuracy	**Prototype-I**	Nill	Nill	Nill
	Prototype-II	60%	55%	52%
	Prototype-III	100%	99%	97%

times before the correct extraction of the required information). Writing accuracy is measured by the number of the letters or characters falsely identified by the Vis-com system or the resident has to repeat himself for the same task. It is observed that prototype-I failed badly and prototype-II performed with an average 55% of accuracy due to illumination, design and resident physiology constraints. However, prototype-III showed the accuracy rate of 98% without any input from the staff member. Thus, the Vis-Com system has minimized the staff member role, as there is no need for a staff member to track the finger movement and identify the letters and then construct a sentence. Naturally, the premise is now only laid for more comprehensive studies on the general usability of the system over longer time.

5 Discussion and Conclusion

In this paper, we have presented our findings of developing a user-centered communication pad. In this case, the user is a cognitive impaired person facing severe challenges in communication due to physical and speech paralysis. To assist the resident and staff member, we devised a computer vision-based hand interactive system to seek enhancing the privacy of the resident in personal communicative matters.
In these types of projects, and as illustrated in the evaluation section, design challenges are easily very diverse and complex due to being rooted in contextual-, technological- and human factors. The system is now ready for more long-term studies as well as investigating how it is possible to derive design guidelines from the many findings during the work on this case study and how these can be applied in new contexts. As the evaluation section demonstrate there are many unforeseen challenges arising from the field. While this is not uncommon in many disciplines it has been visible all along. One example, is that several of the prototypes were well-considered in their technical problem solving. While the hand and finger segmentation was performing

well it did not account for the resident's have a slightly different physiology than expected. Only by confronting the system in the real setting was this possible to fix. And this different phsysiology is highly individually shaped. There are many cases like this, which states two things about this type of work: a: one can not extract all valuable knowledge from the field ahead of development; and, b: a prototype is another constructed reality, which carries its own embedded agendas and must be confronted in situ. Here it stands in a philosophical contrast between Technological Determinism; of what can be constructed to function in ideal cases and that of Social Constructivism, where technology is only meaningful when the user's situation is aligned with implementation. The study thus also illustrates one of the caveats with this type of work - scalability. Custom-fitting technical solutions to individuals is of course a lengthy process. One, arguable, strength is that these types of systems and underlying methodology reflect problem-oriented development, which actually respects the individuality in design and does not assume the user from a generalized (and in some cases steretypic) viewpoint.

In conclusion, we have successfully reduced the size of the communication system and made it portable to be used in almost all scenarios thinkable for the resident (not all other thinkable scenarios). Besides this, the automation of the inferring message system provided convenience to the resident and reduced the staff members involvement, but at the expense of relying on proper light settings as well as accurate hand position for tracking the hand movement over the letter board. However, text recognized is slow due to design constraints such as the slow movement of the resident hand but still faster than analog communication tools used by staff members at the neuro centre.

References

[1] C. M. A. Ilyas, M. A. Haque, M. Rehm, K. Nasrollahi, and T. B. Moeslund, "Facial expression recognition for traumatic brain injured patients," in *International Conference on Computer Vision Theory and Applications*. SCITEPRESS Digital Library, 2018, pp. 522–530.

[2] C. M. A. Ilyas, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2291–2295.

[3] K. Rodil, M. Rehm, and A. L. Krummheuer, "Co-designing social robots with cognitively impaired citizens," in *The 10th Nordic Conference on Human-Computer InteractionNordic Conference on Human-Computer Interaction*. Association for Computing Machinery, 2018.

[4] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[5] A. R. Sarkar, G. Sanyal, and S. Majumder, "Hand gesture recognition systems: a survey," *International Journal of Computer Applications*, vol. 71, no. 15, 2013.

[6] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[7] J. L. Raheja, A. Chaudhary, and K. Singal, "Tracking of fingertips and centers of palm using kinect," in *2011 Third International Conference on Computational Intelligence, Modelling & Simulation*. IEEE, 2011, pp. 248–252.

[8] Y. Li, "Hand gesture recognition using kinect," in *2012 IEEE International Conference on Computer Science and Automation Engineering*. IEEE, 2012, pp. 196–199.

[9] T. Liu, W. Liang, X. Wu, and L. Chen, "Tracking articulated hand underlying graphical model with depth cue," in *2008 Congress on Image and Signal Processing*, vol. 4. IEEE, 2008, pp. 249–253.

[10] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 9–14.

[11] M. Nixon and A. S. Aguado, *Feature extraction and image processing for computer vision*. Academic Press, 2012.

[12] B. Boulay, "Human posture recognition for behaviour understanding," Ph.D. dissertation, Nice, 2007.

[13] E. Kollorz, J. Penne, J. Hornegger, and A. Barke, "Gesture recognition with a time-of-flight camera," *International Journal of Intelligent Systems Technologies and Applications*, vol. 5, no. 3, p. 334, 2008.

[14] A. Bourke, J. O'brien, and G. Lyons, "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm," *Gait & posture*, vol. 26, no. 2, pp. 194–199, 2007.

[15] H. Hasan and S. Abdul-Kareem, "Retracted article: Human–computer interaction using vision-based hand gesture recognition systems: a survey," *Neural Computing and Applications*, vol. 25, no. 2, pp. 251–261, 2014.

[16] Q. Luo, X. Kong, G. Zeng, and J. Fan, "Human action detection via boosted local motion histograms," *Machine Vision and Applications*, vol. 21, no. 3, pp. 377–389, 2010.

[17] M. Alsheakhali, A. Skaik, M. Aldahdouh, and M. Alhelou, "Hand gesture recognition system," *Information & Communication Systems*, vol. 132, 2011.

[18] L. Liu and L. Shao, "Learning discriminative representations from rgb-d video data," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-991-6

AALBORG UNIVERSITY PRESS