Tri-modal Person Re-identification with RGB, Depth and Thermal Features

Møgelmose, Andreas; Bahnsen, Chris; Moeslund, Thomas B.; Clapés, Albert; Escalera, Sergio

Published in:
2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

DOI (link to publication from Publisher):
10.1109/CVPRW.2013.52

Publication date:
2013

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Møgelmose, A., Bahnsen, C., Moeslund, T. B., Clapés, A., & Escalera, S. (2013). Tri-modal Person Re-identification with RGB, Depth and Thermal Features. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): CVPRW 2013* (pp. 301-307). IEEE (Institute of Electrical and Electronics Engineers). https://doi.org/10.1109/CVPRW.2013.52

Tri-modal Person Re-identification with RGB, Depth and Thermal Features

Andreas Møgelmose, Chris Bahnsen, Thomas B. Moeslund
Visual Analysis of People Lab, Aalborg University
am@create.aau.dk

Albert Clapés, Sergio Escalera
Dept. Matemàtica Aplicada i Anàlisis, Universitat de Barcelona, 08007, Barcelona
Computer Vision Center, Campus UAB, 08193, Bellaterra, Barcelona

Abstract

Person re-identification is about recognizing people who have passed by a sensor earlier. Previous work is mainly based on RGB data, but in this work we for the first time present a system where we combine RGB, depth, and thermal data for re-identification purposes. First, from each of the three modalities, we obtain some particular features: from RGB data, we model color information from different regions of the body; from depth data, we compute different soft body biometrics; and from thermal data, we extract local structural information. Then, the three information types are combined in a joined classifier. The tri-modal system is evaluated on a new RGB-D-T dataset, showing successful results in re-identification scenarios.

1. Introduction

Person re-identification is about recognizing people who have passed by a sensor earlier. It is useful in many places where it is desirable to obtain knowledge of the flow of people: airports, transit centers, shopping malls, amusement parks, etc. It can either be knowledge of a single person's movement, or movement patterns in general by combining the patterns of many people. In some cases it is possible to set up a system, which is able to view the entire scene, as in [18, 15]. However, in indoor scenes it is often not feasible to place one camera with a full overview. This is where re-identification enters play. It allows the system designer to place sensors at certain bottlenecks and identify people when they pass these.

Re-identification has the specific distinction from e.g. biometric access control systems that it must be able to enroll new people on-the-fly and without their specific collaboration. On the other hand, the recognition performance does not necessarily have to be as strong as in access control systems, since re-identification systems are more concerned with the general trend of movement as opposed to the movement of each individual.

Re-identification has been an active research area for the past decade, but almost exclusively focused on standard RGB-data. This makes sense since many venues have a large network of already installed RGB surveillance cameras. However, as new and more advanced sensor types become cheaply available, we believe it is time to extend the work to multiple modalities. This is the exact focus of this work, where we present a novel approach that integrates RGB, depth, and thermal data in a re-identification system. An example of RGB, depth, and thermal images for a subject in our dataset is shown in Figure 5.

This paper is structured as follows: Section 2 briefly covers the existing work done on the topic of re-identification, with special focus on the few multi-modal and/or non-RGB-based contributions. Section 3 describes how the inputs from the three modalities are aligned. In sections 4 and 5, the features and re-identification methods are presented. Section 6 shows the dataset and covers the results our system achieves on it. Finally, section 7 concludes the paper.

2. Related work

In [5] soft-biometrics based on RGB data are used to track people across different cameras. Both body and facial soft biometrics are extracted and combined in the final system. The body soft biometrics are all related to color: hair, skin, upper, and lower body clothing. In [6] the notion of tracking people across a multi-camera setup is also followed. Different soft biometric features are reviewed and discussed in the context of re-identification. A part-based appearance approach is found to perform the best, but being sensitive to how the object is divided into parts. In [7] each person is also divided into parts from which features are extracted. The division is here based on finding symmetry axes and the soft biometric features are color histograms, stable color regions and highly structured patches that reoccur. A division is also applied in [9] using similar features. A boosting approach is then introduced to select the most discriminative features. In [1] a similar idea is proposed, i.e., a more reliable classification can be obtained if only the most discriminative features are used for each image region. Moreover they model the uncertainties (covariances) of each feature to improve their results. In [22] a person is divided into six horizontal stripes where each is described in terms of color and texture. The novelty of the work if the formulation of the re-identification problem as a matter of learning the optimal distance measure that minimizes the probability of miss-classification.

All of the above approaches are based on RGB data. Using multi-modal sensing in re-identification is a very new concept and so far only a few works have been reported. In [20] a two-stage recognition approach is followed. First soft-biometrics based on depth data are extracted and secondly RGB data are used in the final classification step. The depth-based soft biometrics are anthropometric measurements and estimated manually. The key finding is that soft biometrics can be used as a pruning step in a recognition system. While this is very interesting, the introduction of manual measurements is not desirable for an automatic re-identification system. In [2] a re-identification method based solely on depth features is presented. The work uses several normalized measures of body parts, calculated from joint positions. Measures of the body's "roundness", which roughly estimates the volume of the torso, are included. High depth resolution is required for this to work and hence it is only suitable when subjects are close to the sensor. The paper is focused solely on the re-identification step and does not treat identification or extraction of joints. In [12] thermal data are used in a re-identification system. The work expands the work reported in [11] where SIFT features are used to model each person. They work on gait data from a side view and can thou track each body part reliably. From each of these a codebook signature is learned over time and combined with the spatial feature distribution found using an Implicit Shape Model.

As opposed to the works described above, in this paper we introduce a truly multi-modal approach based on RGB, depth and thermal data. Moreover, our system is fully automated both in terms of feature extraction, but also when it comes to enrollment.

3. Registration

Since no sensor is able to capture all three modalities at once, a registration of the inputs must take place allowing to map a specific point from one modality to the others. In this work, the Microsoft® Kinect™ for XBOX360 has been used to capture RGB and depth data. A thermal camera (AXIS Q1922) was mounted straight over the Kinect's RGB camera lens with a distance between the lens centers of 70 mm. For registering the tri-modal imagery of this work, we need only to register images from the thermal and visual modalities, as the Kinect provides a factory calibrated registration between the RGB and depth data.

Traditional image registration techniques used for spatially aligning stereo imagery cannot be directly applied to the thermal-visible domain due to the fundamental physical differences of the two modalities, thus rendering the process of finding corresponding features in both imagery is unfeasible. In our setup, objects appear at distances between 1 and 4 meters from the cameras, which makes methods like infinite homography and stereo geometric unusable [13]. Instead we first use stereo rectification to transform the epipolar lines to lines parallel with either the x or y axis [8]. This reduces the search for corresponding points to one dimension. Next we apply the notion that the distance between corresponding points in the two images is inversely proportional to the depth of the points if the cameras are only translated with respect to each other [8]. Since the epipolar lines are transformed to lie along the image scanlines, the disparity between corresponding points will lie mainly either on the x or y axes, and we may thus find the relationship between the inverted depth and the induced disparity and use this property for rectifying the images.

The stereo calibration requires the knowledge of the intrinsic and extrinsic camera parameters of both cameras. In order to determine these, we use the calibration board proposed by [21] with an A3-sized cut-out checkerboard and a heated plate as a viable backdrop. By using standard camera calibration and stereo geometric tools we are able to rectify both images as seen in Figure 1.

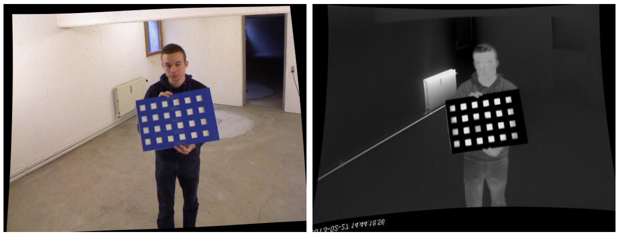

(a) (b)

Figure 1: Stereo rectified multimodal imagery in the (a) RGB and (b) thermal domains.

We used 34 image pairs of the calibration board distributed throughout the entire scene for the calibration of the cameras. For each corner of the chessboard in each image, we extract the corresponding depth. The configuration of cameras placed vertically implies that the disparity of the points in the rectified image lies mainly on the x-axis. Therefore, we use a robust curve fitting tool to find a linear regression that fits the disparity in the x-direction as a function of the inverted distance in the z-direction. The regression is computed off-line for all calibration points and stored for online lookup of the displacement. The result of this procedure is a direct pixel-to-pixel correspondence between the different images.

4. Multi-modal features

The proposed system uses a combination of RGB, depth, and thermal features to perform the re-identification task. This section explains how the feature extraction is performed for each modality. Before the extraction, the subject must first be located at pixel level. The foreground segmentation of the subject is performed on the depth image by means of Random Forest [17]. This process is performed computing random offsets of depth features as follows:

$$f_\theta(\mathcal{D},\mathbf{x}) = \mathcal{D}_{\left(\mathbf{x}+\frac{\mathbf{u}}{\mathcal{D}_\mathbf{x}}\right)} - \mathcal{D}_{\left(\mathbf{x}+\frac{\mathbf{v}}{\mathcal{D}_\mathbf{x}}\right)}, \tag{1}$$

where $\theta = (\mathbf{u},\mathbf{v})$, and $\mathbf{u},\mathbf{v} \in \mathbb{R}^\mathbf{2}$ is a pair of offsets, depth invariant. Thus, each θ determines two new pixels relative to $\mathbf{x}$, the depth difference of which accounts for the value of $f_\theta(\mathcal{D},\mathbf{x})$. Using this set of random depth features, Random Forest is trained for a set of trees, where each tree consists of split and leaf nodes (the root is also a split node). Finally, a final pixel probability of body part membership l_i is obtained as follows:

$$P\left(l_i|\mathcal{D},\mathbf{x}\right) = \frac{1}{\tau}\sum_{j=1}^{\tau} P_j\left(l_i|\mathcal{D},\mathbf{x}\right), \tag{2}$$

where $P\left(l_i|\mathcal{D},\mathbf{x}\right)$ is the PDF stored at the leaf, reached by the pixel for classification $(\mathcal{D},\mathbf{x})$ and traced through the tree j, $j \in \tau$. After this process, the foreground segmentation mask of the subject is transformed to the coordinate system in the two other modalities, and the features are extracted.

The system uses multi-shot person models. Thus, a person is not modeled based on only one frame, but on all frames in a pass. A pass is defined as the act of entering the frame, walking by the camera, and exiting it. In our dataset only one person is present at a time, so no tracking is necessary. Next, we describe how the features from each modality are described and fused in order to perform the on-line re-identification task. Figure 2 summarizes the main modules, modalities and strategies considered in the proposed re-identification system.

4.1. RGB features

After foreground segmentation is performed, the features that are used for the RGB modality are color histograms in two parts, as shown in Figure 3(a). One histogram H_U^{RGB} is derived from the upper body, one H_L^{RGB} from the lower. This is done for each frame in which the subject is detected. A histogram of 20 bins is created for each channel, for a total of 60 bins per body part. Thus, in total the RGB feature

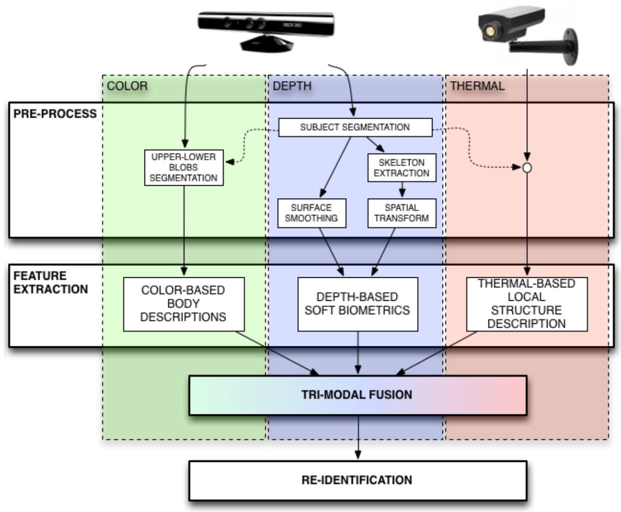
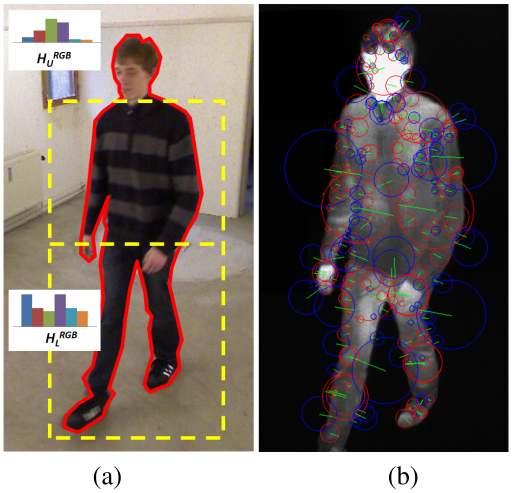


Figure 2: Pipeline of the proposed tri-modal re-identification system.



(a) (b)

Figure 3: (a) Histograms of RGB color distributions for upper body H_U^{RGB} and lower body H_L^{RGB} parts of the subject. (b) Detected SURF keypoints on the thermal modality.

vector has 120 dimensions, and one is created per frame. After a pass ends, the histograms are averaged, and the final feature vector is the mean across the frames.

4.2. Depth features

Given an input depth frame containing a subject (Figure 4(a)), and once the pixel-ground segmentation of the subject into body parts is performed, the skeleton is also extracted applying Mean Shift [17] (Figure 4(b)). Since our dataset contains only raw images, the built-in skeleton-extraction from the Kinect could not be used. Then, the subject point cloud is spatially transformed in order to align the skeleton with the camera frame coordinate system by means of an affine three-dimensional transformation of the point cloud (Figure 4(c)). Note that because of the 3D transformation we loose some information of the body surface due to the lack of information inherent to the viewpoint. Thus, the noisy subject's surface is smoothed (Moving Least Squares surface reconstruction method) and upsampled to fill the holes (Figure 4(d)). Now we can compute soft biometrics from the corrected 3D skeleton and the 3D surface of the aligned body, which can be then inversely transformed to return to the original space and estimate real measurements of the body. From a given depth frame $\mathcal{D}_i$, information invariant to the rotation of the subject with respect to the camera viewpoint can now be extracted. In particular, we have estimated three sets of soft biometrics:

Frontal curve model: The model encodes the distances from the points in subject's surface (transformed and smoothed, as seen in Figure 4(d)) to their corresponding projection line, either head-to-neck or neck-to-torso line. These distances in millimeters are encoded in a real-valued vector $\mathbf{f}_i$, resampled to size 150 and equalized for normalization purposes (Figure 4(e)).

Thoracic geodesic distances: Corresponds to the vector $\mathbf{g}_i$. It contains the length of lines on the body surface from one side of the body to the other. The area in which these are found is the trapezoid defined by left shoulder, right shoulder, right hip, and left hip, and each entry of $\mathbf{g}_i$ contains the geodesic distance in millimeters of a horizontal line in the trapezoid projected to the surface of the torso. $\mathbf{g}_i$ is resampled to size 90 (Figure 4(f)).

Anthropometric relations: Given the extracted body skeleton, the lengths of 7 inter-joint segments connecting the body parts, as shown in Figure 4(c), are computed and stored as $\mathbf{a}_i$.

Thus, the vector representing the set of depth features for a subject in the scene at a particular depth frame $\mathcal{D}_i$ is defined as:

$$\boldsymbol{\delta}_i = \{\mathbf{f}_i, \mathbf{g}_i, \mathbf{a}_i\},$$

where $|\boldsymbol{\delta}_i| = 247$. Finally, the vector describing the subject pass $D = \{F, G, A\}$ is computed by averaging the set of the standardized frame-level depth feature vectors $\{\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_N\}$ as:

$$D = \frac{1}{N} \sum_{j \in N} \frac{\boldsymbol{\delta}_j - \bar{\boldsymbol{\delta}}}{\boldsymbol{\sigma}_{\boldsymbol{\delta}}}, \quad (3)$$

where $|D| = 247$, and $\bar{\boldsymbol{\delta}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\delta}}$ correspond to the mean depth vector and the vector of the standard deviations, respectively. Moreover, as a previous step to this computation and due to the noisy nature of the captured depth data (clothes deformation, waving arms in front of the torso, and so forth), the possible outliers are detected and discarded in each $\boldsymbol{\delta}_i$. This step consists also in standardizing the set of depth feature vectors but to a modified Z-score [10] and discarding those values higher than 3.5 in absolute value.

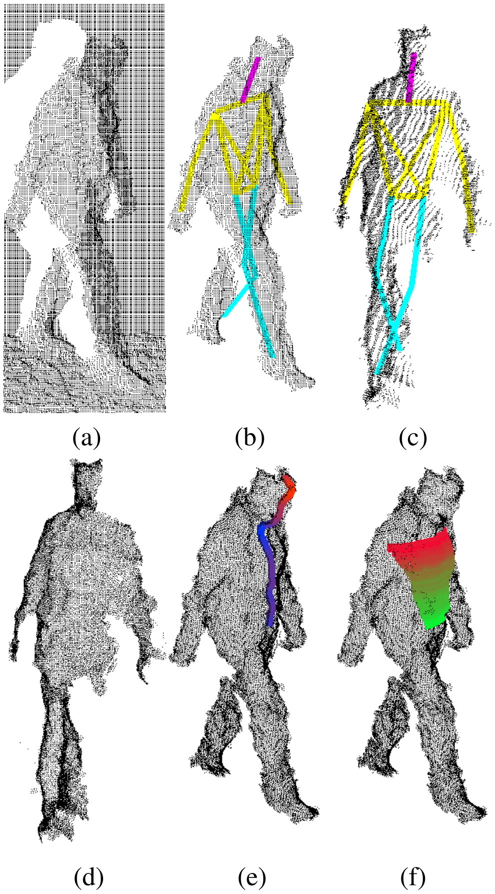


Figure 4: (a) The raw depth data. (b) The pixel-ground segmentation of the subject and the skeleton. (c) After aligning the skeleton with the camera frame. (d) Smoothed data. (e) Vertical projection lines. (f) Geodesic distances.

4.3. Thermal features

Since the thermal images contain no color information, the color histogram approach does not work here. Instead, SURF[3] is employed. Within the contour supplied by the detection stage, SURF-descriptors are extracted. There is no fixed number of descriptors, all that are above a certain quality threshold are extracted. A typical number is around 150 descriptors per subject per frame, depending on the contour's size and quality. As opposed to the RGB histograms there is no direct way to average the descriptors, so the model for people in the thermal modality is all SURF descriptors of the subject extracted over all frames in a pass. We define the set of detected and described SURF points as S, see Figure 3(b).

5. Re-identification

In order to perform the re-identification task, previously computed feature vectors for the three modalities have to be fused and analyzed to classify each subject. The process has two steps:

1. Determine whether the subject is a new or an already known person.
2. Do one of the following two tasks:
   (a) If known, determine the ID of the person.
   (b) If new, enroll the person.

In step 1, a comparison of the current subject with the list of known subjects is done. Taking into account that the set of known persons is built on-the-fly, for the first evaluations only a few comparisons have to be performed.

To estimate whether the subject has to be considered new or re-identified, we compute the following confidence score based on the combination of the three modalities scores:

$$C(U_1, U_2) = \alpha \cdot d_{\text{RGB}}(H_1, H_2) + \beta \cdot \frac{1}{d_{\text{depth}}(D_1, D_2)} + \\ + \gamma \cdot \frac{1}{d_{\text{thermal}}(S_1, S_2)},$$

where $U_1 = \{H_1, D_1, S_1\}$ is the set of three modality descriptors (H_1 color histograms, D_1 depth feature vectors, and S_1 SURF descriptors on the thermal data) for a user in the dataset, and $U_2 = \{H_2, D_2, S_2\}$ are the three sets of descriptors for a new test subject. Coefficients α, β, and γ assigns a proper weight to each of the three modalities scores in a late fusion fashion so that $\alpha + \beta + \gamma = 1$. The weights are static and were set based on experimentation, but for future work, and especially larger datasets, a learning approach for the weights would have to be investigated. The higher the output of $C(U_1, U_2)$, the more reliable re-identification. Because $d_{\text{depth}}(D_1, D_2)$ and $d_{\text{thermal}}(S_1, S_2)$ returns low values in case of good identifications, the reciprocal is used when fused.

For comparing two subjects in the RGB-modality, the Bhattacharyya distance [4] is used:

$$d_{\text{RGB}}(H_1, H_2) = \sqrt{1 - \sum_I \frac{\sqrt{H_1(I) H_2(I)}}{\sqrt{\sum_I H_1(I) \cdot \sum_I H_2(I)}}}, \quad (4)$$

where $d_{\text{RGB}}(H_1, H_2)$ describes the distance between histograms H_1 and H_2, and $H(I)$ is the value of bin I in the histogram H. The distance is a number between 0 and 1, where 0 is a perfect match.

For comparing across subjects in the depth modality $D = \{F, G, A\}$, the following similarity measure is computed:

$$d_{\text{depth}}(D_1, D_2) = W_F(1 - \exp^{-\sum_i w_i (F_1^i - F_2^i)^2}) + \\ + W_G(1 - \exp^{-\sum_j w_j (G_1^j - G_2^j)^2}) + \\ + W_A(1 - \exp^{-\sum_k w_k (A_1^k - A_2^k)^2}). \quad (5)$$

One distance is computed for each of the three depth features, which is in the range [0..1], the lower the distance, the higher the similarity. Coefficients W_F, W_G, and W_A assigns a proper weight to each of the three types of depth feature sets so that $W_F + W_G + W_A = 1$. Moreover, individual feature weights w assign a weight to each particular depth feature value, pre-computed based on a training stage applying ReliefF [16]. In out case the variables were set to $W_F = 0.8$, $W_G = 0.1$, and $W_A = 0.1$.

In the thermal domain, the SURF-descriptors are matched against each other with no spatial information resolved. Each matched feature contributes a vote. Thus the metric is the number of votes for a specific known person across all the frames in the model:

$$d_{\text{thermal}}(S_1, S_2) = \sum_{N_{S_2}} H(n_{\text{votes}}(S_1, S_2)), \quad (6)$$

where $n_{\text{votes}}(S_1, S_2)$ computes the number of matches between SURF descriptors S_1 on the reference image and SURF descriptors S_2 on the test image based on Euclidean distance criterion. H refers to the Heaviside step function, ensuring that each frame in a pass can only contribute one vote, and N are the frames in the model for S_2.

5.1. Determine if new

In order to determine if a person is new, once values for α, β, and γ are established based on a cross-validation of a training stage, two thresholds, T_N and T_R are also experimentally computed. If $C < T_N$, the subject is considered new. If $C > T_R$ the subject is assigned a known ID (re-identified). Since a false positive is more serious than a false negative in re-identification, we have a buffer zone when $T_N \leq C \leq T_R$ where the system ignores the subject because we are uncertain whether it is a new person or just a bad match to an existing one. In our system we used $T_N = 6$ and $T_R = 10$, but the exact value of the thresholds seemed to be relatively flexible.

5.2. ID determination

The assignment of an ID to an already existing user for re-identification is straightforward using the confidence score C obtained from the previous step. If the user has been determined as already known, it means that the majority of votes are given to a particular user ID which is the one assigned in the re-identification task.

6. Evaluation

Several re-identification datasets with RGB [9, 14] and RGB-D data [2] exist, but to the best of our knowledge no dataset containing all three modalities exists. We have therefore recorded a novel re-identification tri-modal dataset.

The dataset consists of 35 people passing by the sensors twice for 70 passes in total. The vantage point is up and

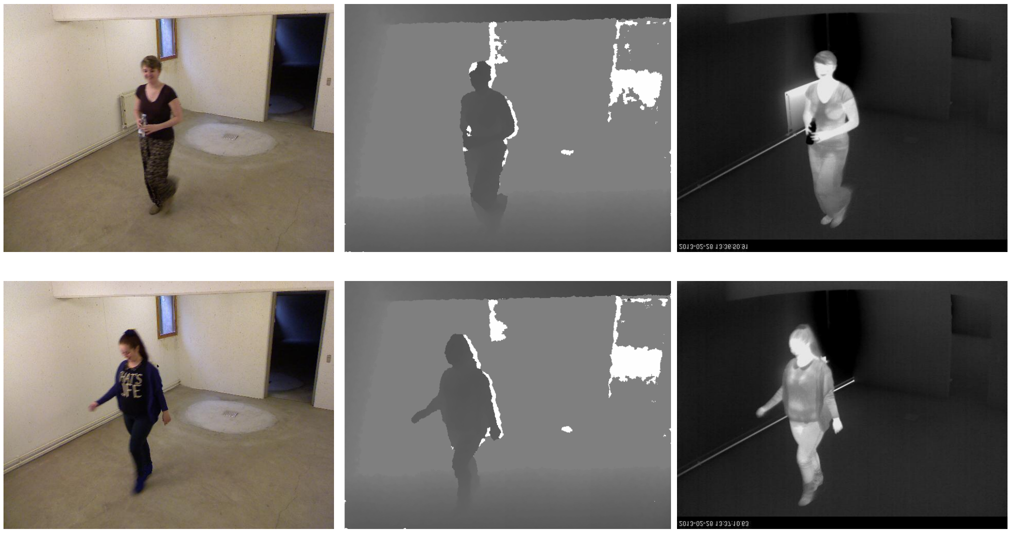


Figure 5: Sample images from the tri-modal dataset. left, middle, and right are RGB, depth, and thermal, respectively.

slightly off to the side to mimic a classic surveillance camera setup. All images are 640×480 pixel. Some sample images from each modality are shown in Figure 5.

The tests were conducted by first extracting the aforementioned features from all passes. As this system is a re-identification system with online enrollment, there is no explicit training phase. Instead, the persons are enrolled if they are very different from previous seen persons.

Since the order of passing will influence the re-identification performance, the system was tested in a random 5-cross validation. We tried the different sets of modalities as input features and found that the best combination of features is the late fusion considering the three sets of modalities with weights: $\alpha = \frac{1}{3}$, $\beta = \frac{1}{3}$, and $\gamma = \frac{1}{3}$ to fit the tri-modal scheme. The results are presented both individually and averaged in terms of: A) passes correctly classified as a new person, B) passes wrongly classified as a new person, C) the number of correctly re-identified persons, D) the number of wrongly re-identified persons, and E) the number of persons ignored, see Table 1.

If an application requires every single person to be re-identified, then it can be inferred from the table that the performance of our system is 39.4%. In most cases, however, re-identification is used to measure the overall flow and the important issue is therefore to have an acceptable number of true positives and a low number of false positives, where especially the latter is clearly obtained in our system. For comparison a commercial re-identification system based on Wi-Fi signals from smartphones operates with a performance of approximately 50% [19].

Similar to others working on re-identification we also compute the CMC-curve to show the recognition performance for different rank values, see Figure 6. Each of the dashed lines is a CMC-curve for a single run. The thick black line is the mean CMC of the 5 runs.

	A	B	C	D	E
Run 1	35	10	16	0	9
Run 2	34	12	12	1	11
Run 3	33	13	13	1	10
Run 4	34	12	15	1	8
Run 5	34	10	13	2	11
Average	34	11.4	13.8	1	11
Percentage			93.2%	6.8%	

Table 1: Re-identification results.

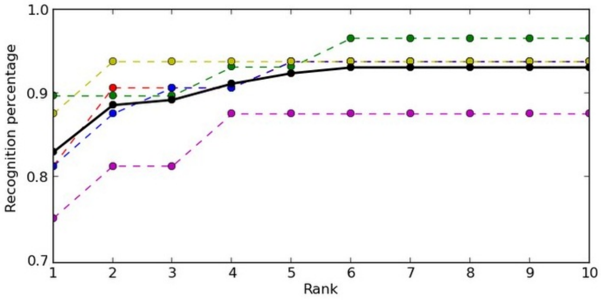


Figure 6: CMC-curve performance.

Since this is the first work on tri-modal re-identification we cannot compare our results directly with those of others. Instead in Table 2 we list the rank-1 results of previous works. Please note that very different datasets and setting were used in these works and that no final conclusions therefore can be drawn. The results, however, seem to indicate the quality of our tri-modal approach, especially since we do not have a training phase as most others do.

Work	[1]	[2]	[5]	[6]	[7]	[9]	[12]	[20]	[22]	Our
Data	RGB	Depth	RGB	RGB	RGB	RGB	Thermal	RGB-D	RGB	RGB-D-T
Rank-1	51%	12%	N/A	82%	67%	43%	98%	78%	26%	82%

Table 2: Data types and rank-1 results of recent re-identification works. Note that several works test on a number of different settings and different datasets. In such cases the table contains the average of the best results.

7. Concluding remarks

We proposed a tri-modal re-identification system based on RGB, depth, and thermal descriptors. Three modalities were aligned, and robust discriminative features codifying soft biometrics were computed. The modalities were combined in a late fusion fashion, being able to predict a new user in the scene as well as to recognize previous users based on a combined rule cost. We tested our tri-modal re-identification system on anovel tri-modal dataset. Our results showed that the combination of all three modalities is the one that achieved better performance. A place to improve the system is in the determination of new persons. Nearly all new persons are detected as such, but there is a substantial amount of wrong New Persons. That is not a big issue with regards to re-identification performance, as presumably they will also be difficult to re-identify (they are only detected as new because they are not similar to the known persons), and in many applications it is not critical to be able to re-identify each and every subject. However, fewer wrong New Persons will result in a lower absolute re-identification rate.

References

[1] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to Match Appearances by Correlations in a Covariance Metric Space. In *ECCV (3)*, volume 7574 of *LNCS*, pages 806–820. Springer, 2012. 2, 7

[2] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with RGB-D Sensors. In *ECCV Workshops (1)*, volume 7583 of *LNCS*, pages 433–442. Springer, 2012. 2, 5, 7

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 4

[4] G. Bradski and A. Kaehler. *Learning OpenCV*, chapter 7, pages 201–202. O'Reilly, 2008. 5

[5] M. Demirkus, K. Garg, and S. Guler. Automated person categorization for video surveillance using soft biometrics. pages 76670P–76670P–12, 2010. 1, 7

[6] G. Doretto, T. Sebastian, P. H. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *J. Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011. 1, 7

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 7

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2000. 2

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-identification by Descriptive and Discriminative Classification. In *SCIA*, volume 6688 of *Lecture Notes in Computer Science*, pages 91–102. Springer, 2011. 2, 5, 7

[10] B. Iglewicz and D. Hoaglin. *How to Detect and Handle Outliers*. ASQC basic references in quality control. ASQC Quality Press, 1993. 4

[11] K. Jüngling and M. Arens. Local Feature Based Person Reidentification in Infrared Image Sequences. In *AVSS*, pages 448–455. IEEE Computer Society, 2010. 2

[12] K. Jüngling and M. Arens. A multi-staged system for efficient visual person reidenticationl. In *Conference on Machine Vision Applications, Nara, Japan*, 2011. 2, 7

[13] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(2):270–287, 2007. 2

[14] C. C. Loy, T. Xiang, and S. Gong. Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding. *Int. J. Comput. Vision*, 90(1):106–129, Oct. 2010. 5

[15] B. E. Moore, S. Ali, R. Mehran, and M. Shah. Visual crowd surveillance through a hydrodynamics lens. *Commun. ACM*, 54(12):64–73, 2011. 1

[16] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning. *Machine Learning*, 53:23–69, 2003. 5

[17] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304. IEEE, 2011. 3

[18] B. Solmaz, B. E. Moore, and M. Shah. Identifying Behaviors in Crowd Scenes Using Stability Analysis for Dynamical Systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):2064–2070, 2012. 1

[19] B. Systems. Urban Planning. http://www.bliptrack.com/urban/products/bliptracktm-sensor/, 2013. 6

[20] C. Velardo and J. Dugelay. Improving Identification by Pruning: A Case Study on Face Recognition and Body Soft Biometric. In *WIAMIS*, pages 1–4. IEEE, 2012. 2, 7

[21] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark. A mask-based approach for the geometric calibration of thermal-infrared cameras. *Instrumentation and Measurement, IEEE Transactions on*, 61(6):1625–1635, 2012. 2

[22] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656. IEEE, 2011. 2, 7