Multimodal Person Re-identification Using RGB-D Sensors and a Transient Identification Database

Møgelmose, Andreas; Moeslund, Thomas B.; Nasrollahi, Kamal

Published in:
International Workshop on Biometrics and Forensics

DOI (link to publication from Publisher):
10.1109/IWBF.2013.6547322

Publication date:
2013

Document Version
Early version, also known as pre-print

MULTIMODAL PERSON RE-IDENTIFICATION USING RGB-D SENSORS AND A TRANSIENT IDENTIFICATION DATABASE

A. Møgelmose, T. B. Moeslund, and K. Nasrollahi, Aalborg University

ABSTRACT

This paper describes a system for person re-identification using RGB-D sensors. The system covers the full flow, from detection of subjects, over contour extraction, to re-identification using soft biometrics. The biometrics in question are part-based color histograms and the subjects height. Subjects are added to a transient database and re-identified based on the distance between recorded biometrics and the currently measured metrics. The system works on live video and requires no collaboration from the subjects. The system achieves a 68% re-identification rate with no wrong re-identifications, a result that compares favorable with commercial systems as well as other very recent multimodal re-identification systems.

Index Terms— Re-identification, RGB-D, multimodal

1 INTRODUCTION

Person re-identification is useful in many contexts, and can be used as a forensics tool in most situations where surveillance cameras has captured an incident. Re-identification is the act of recognizing persons entering a camera's field of view and have been seen previously by a different camera, or by the same camera at a different time instance. The crucial difference between this and tracking is that for re-identification there is expected to be a significant spatial or temporal difference between observations, making it impossible to rely on simple motion dynamics as tracking often does. Instead, soft biometrics are used to decide if a subject has been seen before.

A number of challenges and characteristics set re-identification apart from traditional tracking and hard-biometric recognition:

- The set of re-identifiable persons must be updated on the fly; there can be no enrollment phase that requires direct participation from the subjects.
- There is no – or only weak – constraints on the pose of subjects, so the system must be robust to pose changes.
- Persons must be re-identifiable at distances where sensor resolution is generally not sufficient for traditional face recognition.
- The database containing the subjects has a transient nature since subjects are generally not relevant if they have not been re-identified after a certain time span – then they have probably left the area.

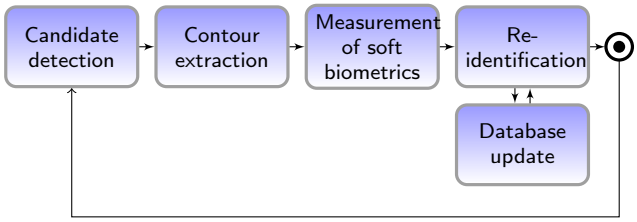

Figure 1: The flow of the method discussed in this paper. The loop runs once per frame.

Some applications of re-identification does not require all recorded persons to be re-identified. An example is the commercial system from Blip Systems [9], which does person flow tracking in airports based on radio signatures from mobile phones. It has a re-identification rate of around 10%, which is sufficient for a representative flow map.

Because the re-identification scenario can be harder than traditional recognition due to the worse data quality, it is an obvious idea to use more sensor modalities. With the advent of the Microsoft Kinect and similar structured light-based sensors (ASUS Xtion and the PrimeSense Sensor), RGB-D sensors have become much more accessible and affordable, and using them in larger surveillance applications does not seem impossible. While sensors relying on structured light have some issues, especially with outdoor use, we believe that in the near future, many more modalities – such as depth – will be integrated in surveillance cameras, and as such it is prudent for the surveillance computer vision research community to direct its attention toward multimodal methods.

The main contribution of this paper is a RGB-D based re-identification system. It performs all the steps necessary in these kinds of systems: Person detection, measurement of soft biometrics, forming and maintaining a transient database of subjects, and re-identifying subjects, whereas previous RGB-D based contributions (see section 2) has only covered parts of the process.

The rest of the paper is structured as follows: Section 2 takes a brief look at related work in the area of re-identification. Section 3 describes the structure of the re-identification methods used in the system, and contains subsections going in further details with each step. The transient database is treated in section 4, which is followed by exper-

978-1-4673-4989-5/13/$31.00

iments and tests of the system in section 5. The closing remarks can be found in section 6.

2 RELATED WORK

Re-identification is a relatively young field, and most contributions so far are based on regular camera input. Notable examples include [1, 12, 5, 11, 8].

Re-identification using RGB-D sensors is still in its infancy; only a few papers on RGB-D re-identification exist. [2] present a re-identification method based solely on depth-features using several normalized measures of body parts, calculated from joint positions. They include measures of the body's "roundness", which can act as a crude proxy for volume. This, however requires a high depth resolution, and 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, while our paper presents a full system. Another approach for re-identification using soft biometrics was put forth in [10], but they use manual measurements instead of automatic analysis of RGB-D images.

3 METHOD OVERVIEW

This system covers all stages through a complete re-identification flow. An overview is presented in figure 1. First thing to happen is the candidate detection: Before any re-identification can take place, it is necessary to know if - and where - the person is. Next step is contour extraction. A detector usually only returns a bounding box, and possibly even a bounding box that does not fit closely around the person. When extracting the biometrics, it is important not to extract information from the background, but from the subject alone, since the subject will appear on different backgrounds later and must be re-identifiable then. Next step is measurement of the desired soft biometrics to generate a descriptor for this particular subject. Finally, the candidate is either identified, added to a database of previously seen persons, or ignored due to too low data quality (this particular situation is covered in further depth in section 4).

3.1. Candidate detection

The detection stage consists of a state-of-the-art HOG-SVM detector trained on the INRIA dataset [4] run on the RGB image.

Since re-identification is not tracking, continuous detection in every frame is not necessary. As long as one good detection - or however many the re-identification process takes - is present, the re-identification can be performed. In the context of flow analysis, it is also enough to detect and re-identify the 10% of subjects with the strongest response (as done by [9]). Because of this, the detection rate can be lowered, to the benefit of the false detection rate.

Figure 2: A real subject split in parts. The histograms are calculated based on pixel values inside the shaded areas in the boxes. RGB image on the left, depth image on the right.

3.2. Contour extraction

The depth image is used for contour extraction. When seen from a distance, persons in the depth image are generally a plane with only small depth variations, so a flood fill is used. A seed point is selected in the middle of the chest of the detected person. The starting point is defined as

$$(x, y) = \left(x_b + \frac{w_b}{2}, y_b + \frac{h_b}{3}\right) \quad (1)$$

where x_b and y_b are the coordinates of the upper left corner of the detection bounding box, and w_b and h_b are the width and height, respectively.

A problem arises when the fill reaches the floor. A line on the floor will be in the same depth as the feet, and thus the fill continues onto the floor. To counter this, during the initialization of the system, the ground plane is estimated. This is done by manually clicking a number of points that are on the floor and fitting a plane through these using a least squares fit performed by SVD factorization[7]. Pixels in the depth image that are near the estimated ground plane are discarded from the fill.

3.3. Measurement of soft biometrics

Two soft biometrics are used in the system: A part-based color histogram and the subject's height. The height is found by subtracting the y-values in world-coordinates for the up most and the lowest point in a contour.

The histogram is calculated on parts of the subject to account for the differing colors between leg garments and jacket/shirt. According to [6], the legs occupy 0% to 55% of the full body height and the torso from 55% to 84% of the full height. This implementation takes its base in these figures, but since the division between legs and torso can vary from person to person, an undefined zone is introduced at the middle of the body, which is not counted in either histogram. Thus, the division used here can be seen in fig. 2.

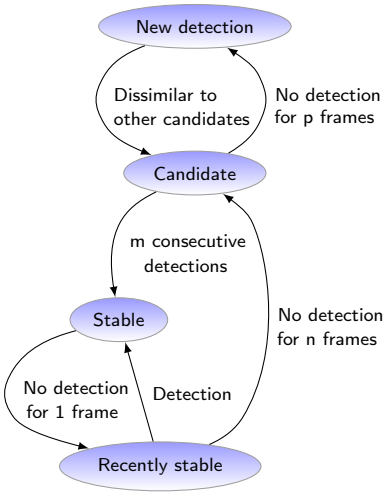

Figure 3: State diagram for the transient database.

For each part, a histogram is created for each of the R, G, and B channels with 20 bins. These are concatenated for at total of 60 bins per part. Then the part histograms are also concatenated and the full 120 bin histogram is normalized so all bins sum to 1. This makes sure that comparison across different sizes is possible, and evens out changes in lighting.

Because this system runs on real-world data and the segmentation becomes unstable at far distances, only subjects within 4 m of the camera are considered.

3.4. Re-identification

The re-identification step consists of a comparison of the candidate to the persons saved in the transient database. The details on the database can be found in the next section, but the most basic functionality is comparison of the histograms. This is done using the Bhattachariyya distance [3]:

$$d(H_1, H_2) = \sqrt{1 - \sum_I \frac{\sqrt{H_1(I)H_2(I)}}{\sqrt{\sum_I H_1(I) \cdot \sum_I H_2(I)}}} \quad (2)$$

where $d(H_1, H_2)$ is the distance between the histograms H_1 and H_2, $\bar{H}$ denotes the norm of a histogram, N is the number of bins, 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.

The next section describes how the height and the distance between histograms are used with the database.

4 TRANSIENT DATABASE

The purpose of the transient database is to contain the previously detected persons. Because subjects are often only seen briefly, there are very few samples per person and they are often not very structured. This makes the creation of a parametric model unreliable. Instead, we model each person with all the previous heights and histograms that have been connected to her. This gives a broad model of each candidate in many poses, orientations, and sizes. A person is then re-identified if the query biometrics are sufficiently close to either of the existing samples in a database entry for one person. In essence this is a box classifier with the histogram distance and the height as features. A box classifier makes sense since it is a reasonable assumption that the height and the clothing color are not correlated.

To minimize noise, some temporal constraints must be fulfilled before a person is considered as recognized. Thus, the detected subjects can have several states (see fig. 3):

- New detection
- Candidate
- Stable
- Recently stable

New detection is the initial state for any new detection. If the subject is sufficiently dissimilar to the existing database entries, she is added to the database as a *candidate*. Because there is always a risk of false detections, the candidate must have been detected for at least m consecutive frames in order to become *stable*. Only stable persons are considered for re-identification.

If a stable person is not successfully re-identified, she is transferred to *recently stable*, which allows her to regain stable-state from just a single detection. If a person has been recently stable for n frames, she is transferred to the candidate stage and must regain stability on equal terms with all other candidates. Finally, a candidate that has not been seen for a given time measured in frames, p, is likely to have left the venue, and is thus discarded completely from the database. This ensures that the database size remains at a size where it is feasible to search for new candidates quickly.

5 EXPERIMENTS

A set of recordings was used to evaluate the system. They contain 25 subjects which walk past the camera twice each (not in any particular order). In total the test set consists of 7800 frames. The results are presented in table 1. They were counted on each pass of a subject, so one pass with a correct re-identification counts for 1 in that category.

Correct re-identifications are exactly that: The subject is identified with a correct label. Ambiguous re-identifications are instances where the subject is re-identified as several people during a pass, but at least one of them is the correct label. They are still only detected as a single person, but the identification differs from frame to frame. Not enrolled means that a person is not re-identified due to the fact that the first pass did not result in any sufficiently good features, so she was never enrolled in the database. Not re-identified means that a person was enrolled, but not recognized in the subsequent pass. Finally, wrong re-identification describes the case where a person is erroneously classified with the label of an-

Table 1: Experimental results

Subjects	Absolute	%
Correct re-identifications	17	68%
Ambiguous re-identifications	2	8%
Not enrolled	5	20%
Not re-identified	1	4%
Wrong re-identifications	0	0%

other person.

76% of the subjects are correctly identified (albeit with 8% ambiguously re-identified), 20% was not identified due to missing enrollment on their first pass, and a single person was not re-identified, even though she was correctly enrolled. A design choice was to ensure that wrong re-identifications would not occur. This has been achieved successfully, but it of course has an adverse effect on the re-identification percentage. It should be noted, however, that a successful re-identification of 76% is significantly better than what [9] are capable of for their commercial system.

A very recent RGB-D re-identification study by Barbosa et. al. [2] achieves a rank 1 re-identification rate of around 15%, significantly worse than our results. In this context, rank relates to the confidence with which the person is re-identified. When a subject is re-identified, the system makes a ranked list of likely labels. Rank 1 means that the correct label is the highest ranked. Rank 5 would mean that the correct label is within the 5 highest ranked labels. Thus, their rank 1 result of around 15% is the number that can be directly compared with our results of 76%. They report their main results as the normalized area under the Cumulated Match Characteristics (CMC) curve where the x-axis is the rank and the y-axis is re-identification rate. Since our system is intended for fully automated use, anything but the rank 1 result is largely irrelevant, but if the system were to be used in a supervised context, the nAUC is indeed interesting.

6 CONCLUDING REMARKS

This paper presented a full RGB-D based person re-identification system. On our test set, it achieves a re-identification rate of 68%. This outperforms both commercial systems for person flow tracking [9] and very recent multimodal systems [2]. It works on real-world data and covers the full system from detection through contour extraction, measurement of soft biometrics and the actual re-identification. The system exhibits a weakness with similarly dressed persons, something that might be solved be adding more advanced biometrics. Tracking might improve performance. Moreover, like others, the system assumes no occlusions, something that should be addressed in future work.

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] 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.

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

[4] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

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

[6] P. Fihl, R. Corlin, S. Park, T. B. Moeslund, and M. M. Trivedi. Tracking of Individuals in Very Long Video Sequences. In *ISVC (1)*, volume 4291 of *LNCS*, pages 60–69. Springer, 2006.

[7] G.H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420, 1970.

[8] 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.

[9] Blip Systems. Blip Track Airport. http://www.bliptrack.com/airport/area-of-operations/, 2012.

[10] 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.

[11] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and Appearance Context Modeling. In *ICCV*, pages 1–8. IEEE, 2007.

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