FACE- AND APPEARANCE-BASED PERSON IDENTIFICATION FOR FORENSIC ANALYSIS OF SURVEILLANCE VIDEOS

Christian Herrmann[1], Jürgen Metzler[2], Dieter Willersinn[3] and Jürgen Beyerer[4]

[1]*christian.herrmann@iosb.fraunhofer.de,* [2]*juergen.metzler@iosb.fraunhofer.de,*
[3]*dieter.willersinn@iosb.fraunhofer.de,* [4]*juergen.beyerer@iosb.fraunhofer.de*
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB,
Fraunhoferstr. 1, 76131 Karlsruhe (Germany)

1 INTRODUCTION

The increasing availability of surveillance cameras is both an opportunity and a challenge for forensic crime investigation. For serious crimes, video footage offers a widely accepted opportunity to identify criminals and reconstruct the events to find further offenders. Because this is often a highly manual task, automated video analysis methods are welcome to efficiently handle the growing amounts of video data. The research project MisPel (Multi-Biometriebasierte Forensische Personensuche in Lichtbild- und Videomassendaten) funded by the German Ministry of Education and Research addressed this field by creating and combining several automated video analysis tools into a demonstration system. A forensic analysis system is always controlled by a human operator who selects which data are to be analyzed and what kind of analysis should be performed. Here, the focus will be on one specific part: the identification of persons to find all occurrences of an offender in the video data. Assuming that an offender has caught the operator's attention, the aim is to assist the operator by finding further occurrences of this particular person in the relevant video data. This contribution focuses on the technical part of the data extraction.

2 FACE RECOGNITION

The first of the two presented methods to search for persons is via face recognition. This method usually works by extracting several features from a detected face which are consequently used for the comparison to further face representations. As result a ranked list of face matches is created showing the best match at the top. The major challenge when analyzing surveillance videos is their quality. Large capture distances lead to low resolution, movement of a person to motion blur, small data connections to severe compression artifacts and bad illumination to noise in the captured image (compare Fig. 1). By adjusting the collected features and fusing the sequential information of consecutive frames of the videos, some of the effects can be compensated [1]. Further benefits for low resolution data are possible by super-resolution including an extension to non-frontal head poses which are the usual case in surveillance data [2]. As a rule of thumb, one can say that a minimum face size of about 20-30 pixels is required to successfully perform face recognition. If the face size drops below that range, it is already the detection of the faces in the video that struggles significantly. An experiment with the widely used Viola-Jones-based face detector [3] shows that when trying to detect faces again on a lower resolution, that were detectable on a high resolution of 100 pixels, the performance degrades heavily below 20 pixels as is shown by Table 1. This indicates that the images need to have a certain quality and resolution to enable successful face recognition. Despite the possibilities to enhance low quality face recognition, there still exist a lot of surveillance data, where this approach is infeasible. Not only if faces are too small, but also if a person is only visible from behind or the face is occluded.

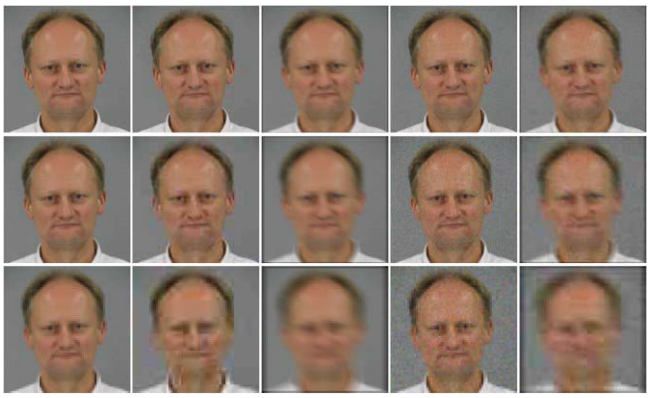

Fig. 1. Different effects that degrade the face quality. From left to right: resolution, compression artifacts, motion blur, noise and a combination of all. Severity of effect increases from top to bottom.

Table 1. Face detection rate depending on face resolution

face size (in pixels)	detection rate
100	100 %
30	98 %
25	95 %
20	90 %
15	72 %
12	35 %

3 APPEARANCE-BASED PERSON RE-IDENTIFICATION

Compared to face recognition, appearance-based person re-identification by the body relies mostly on the patterns and colors of a person's clothes. Because these cover a larger area on a person's body than a face does, this strategy is more suitable for lower image resolutions. However, a person can easily change the appearance, for example, by changing clothes, so this method might fail in crime investigations involving video data from several different days. Appearance-based person re-identification in cameras has become an intensive research topic in recent years [4]. Given an image region of a person (probe), the objective is to re-identify it in further video data (gallery). The result is a ranking of all comparisons sorted by their similarities which enables a human operator to quickly re-identify persons between video clips. The system chart in Fig. 2 shows an overview of the entire identification approach including the appearance-based person re-identification that is based on the work in [5] and [6].

Here, persons are represented by covariance descriptors that were introduced by Porikli et al. ([7]). They represent an image region by a covariance matrix of image features which is a natural way of fusing multiple features and, as shown in [7] and [8], there are several advantages of using covariance descriptors: they are invariant to mean changes, e.g. invariant to identical shifting of color values, insensitive to noise, support scale invariant features and offer an efficient fusion of multiple features. Furthermore, an evaluation of several tracking methods showed that the covariance descriptors-based tracking is the most appropriate one for crowd and riot scenarios [10] which encouraged us to use them for appearance-based re-identification ([5],[6]).

Let R_1 be an image region. First, for each pixel inside R_1 features are computed. Here, we use the x- and y-coordinates of the image pixels and the color values R, G and B. Then, d-dimensional feature vectors are constructed – one for each pixel inside R_1. Let $\{f_i\}_{i=1\dots n}$ be a set of feature vectors of the W-width and H-height rectangular R_1 and $f_i = (x, y, R(x,y), G(x,y), B(x,y))^T$ a feature vector at the pixel with the coordinates (x, y) and color values $R(x,y), G(x,y), B(x,y)$. Then the covariance, using the mean-vector μ_{R_1} of $\{f_i\}_{i=1\dots n}$, is $Cov_{R_1} = \frac{1}{WH}\sum_{i=1}^{WH}(f_i - \mu_{R_1})(f_i - \mu_{R_1})^T$. An efficient way to compute the covariance matrices for image regions of arbitrary size can be found in [10]. In order to compare two covariance descriptors, a non-Euclidean metric is required since covariance matrices do not reside in a vector space. The set of positive definite symmetric matrices can be formulated as a Riemannian manifold as described in [11]. With their proposed Riemannian metric, the (geodesic) distance between two given covariance matrices $\sum_1$ and $\sum_2$ is $d(\sum_1, \sum_2) = \sqrt{< log_{\sum_1}(\sum_2), log_{\sum_1}(\sum_2) >_{\sum_1}}$, where

$log_{\sum_1}(\sum_2) = \sum_1^{\frac{1}{2}} \log(\sum_1^{-\frac{1}{2}} \sum_2 \sum_1^{-\frac{1}{2}}) \sum_1^{\frac{1}{2}}$. The Riemannian metric is used for the person matching that is based on the output of a person tracker. As first step of the re-identification process, for each image region in the sequence of the person of interest (probe) a covariance descriptor is calculated as described above. Then, they are matched with the covariance descriptors of the gallery, resulting in a ranking of the gallery sequences.

4 COMBINATION

Based on the respective characteristics, the domain of appearance-based re-identification is the short time analysis, while face recognition helps for an analysis involving larger time periods. Combining both methods offers a broader scope of possible applications for the automated analysis.

Each method creates a ranking of the further occurrences of the person of interest. Both rankings might partially overlap in some results if both methods are able to detect the same occurrence. Due to information about timestamp and image location, they can easily be matched and fused. This prevents that the operator has to watch the same result twice.

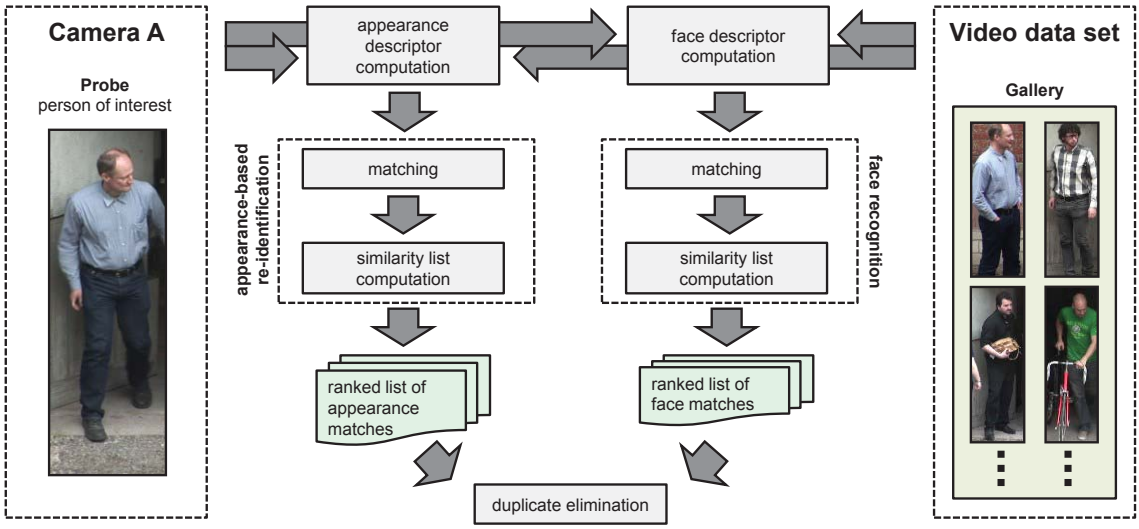

Fig. 2. System chart of the combined system.

5 RESULTS AND CONCLUSION

Applying the strategy to a set of surveillance videos leads to the results shown by Fig. 3. Using an example query person, the appearance-based and face-based searches lead to two different rankings. After finding the duplicates, one can see the benefits of the method combination. The best appearance-based match was missed by the face matching, because the person is only visible from behind. On the other hand, the face recognition finds two further occurrences that were impossible to detect for the appearance-based strategy because of different clothing and occlusion of the body. The results show that the pursued combination strategy leads to beneficial results compared to using each method alone.

ACKNOWLEDGEMENT

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as part of the MisPel program under grant no. 13N12062 and 13N12063.

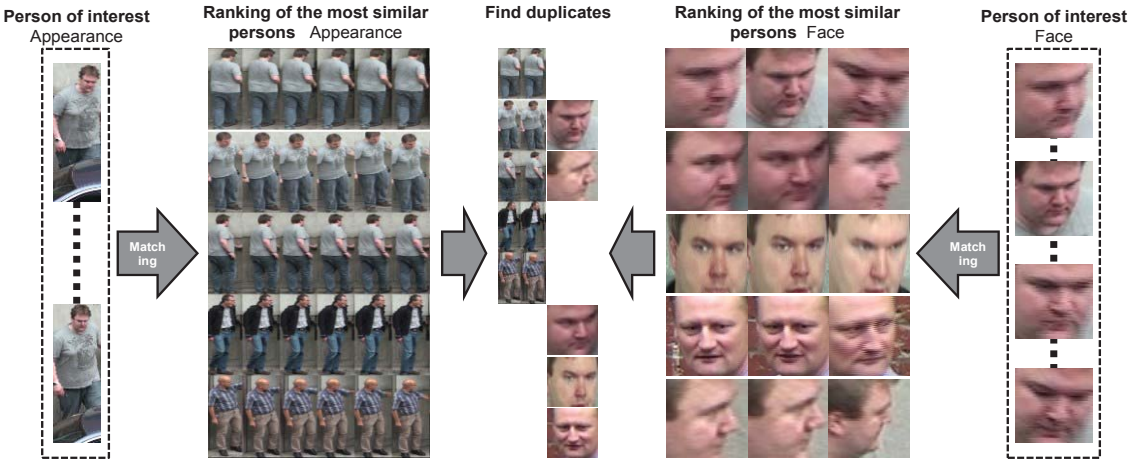

Fig. 3. Matching results including individual and fused rankings of the video sequences, each starting with the gallery sequence that is most similar to the probe.

REFERENCES

[1] Herrmann, C. (2013). *Extending a local matching face recognition approach to low-resolution video*. In Advanced Video and Signal Based Surveillance, pp.460-465.

[2] Qu, C.; Herrmann C.; Monari, E.; Schuchert, T. and Beyerer J. (2015). *3D vs. 2D: On the Importance of Registration for Hallucinating Faces under Unconstrained Poses*. In Computer and Robot Vision.

[3] Viola, P. and Jones, M. (2001). *Rapid Object Detection Using a Boosted Cascade of Simple Features*. In Computer Vision and Pattern Recognition.

[4] Bedagkar-Gala, A. and Shah, S. K. (2014). *A survey of approaches and trends in person re-identification,* ELSEVIER Journal, Image and Vision Computing, vol. 32, pp. 270-286.

[5] Metzler, J. (2012). *Appearance-based re-identification of humans in low-resolution videos using means of covariance descriptors*. In Advanced Video and Signal Based Surveillance, pp. 191-196.

[6] Metzler, J. (2012). *Two-stage appearance-based re-identification of humans in low-resolution videos*. In Workshop on Information Forensics and Security, pp. 19-24.

[7] Tuzel, O.; Porikli, F. and Meer, P. (2006). *Region covariance: A fast descriptor for detection and classification*. In European Conference on Computer Vision, vol. 2, pp. 589-600.

[8] Porikli, F.; Tuzel, O. and Meer, P. (2005). *Covariance Tracking using Model* Update Based on Means on Riemannian Manifolds. In Computer Vision and Pattern Recognition, vol. 1, pp. 728-735.

[9] Hübner, Y.; Metzler, J.; Dürr B.; Jäger U.; Willersinn D. (2008). *Assessment and optimization of methods for tracking people in riot control scenarios*. In Proc. SPIE 7114, pp. 191-196.

[10] Porikli, F. and Tuzel, O. (2006). *Fast Construction of Covariance Matrices for* Arbitrary Size Image Windows. In International Conference on Image Processing, pp. 1581-1584.

[11] Pennec, X.; Fillard, P. and Ayache, N. (2006). *A Riemannian Framework for Tensor Computing*. International Journal of Computer Vision, vol. 66, pp. 41-66.