Person re-identification in multi-camera networks

Kai Jüngling, Christoph Bodensteiner, Michael Arens
Fraunhofer IOSB
Ettlingen, Germany
{kai.juengling,christoph.bodensteiner,michael.arens}@iosb.fraunhofer.de

Abstract

In this paper, we present an approach for person re-identification in multi-camera networks. This approach employs the Implicit Shape Model and SIFT features for person re-identification. One important property of the re-identification approach is that it is closely coupled to a person detection and tracking and uses SIFT feature models which are built during the tracking. We hold this coupling to be an important point because re-identification depends on models that are to be acquired during tracking. These models are then used to re-identify a person when it reappears in the system's field of view. Re-identification itself is performed in a 3-staged approach which allows for efficient re-identification and is perfectly suited for distributed processing where bandwidth concerns are relevant. We show that this re-identification approach – which was formerly only evaluated for single camera person re-identification can be successfully applied to the task of multi-camera re-identification. Evaluation in a challenging real-world multi-camera scenario shows that the generic approach which does not use color or other sensor specific features and thus is applicable independently of such sensor specifics – shows performance at least comparable to specialized state-of-the-art approaches.

1. Introduction

Object, and more specifically person tracking is an indispensable part of many of today's computer vision applications. In many cases where high-level video analysis is necessary, specifically in areas like visual surveillance, it is not sufficient to track a person while it continuously appears in the field of view of a single camera, but to re-identify a person after it has left the system's field of view and reenters it again. The system's field of view hereby can refer to a single or multiple cameras. This re-identification is essential in many applications like multi-camera tracking. Even when referring to a system with a single camera, re-identification can be necessary, e.g. to determine if a person visits a shop-window multiple times, to check if the same person or another person picks up a bag that someone has left before, and also to detect suspicious behavior which can be constituted by visiting the same place multiple times.

In this paper, we approach the problem of person re-identification. We build on the re-identification approach described in [8] and revised in [10]. The essential of this approach is that it builds on the Implicit Shape Model (ISM) [11] and SIFT [12] features for both person tracking and re-identification. The re-identification uses the SIFT features and ISM characteristics collected online during the tracking of a person to re-identify this person later on. The overall approach thus has some major advantages over current re-identification approaches:

(i) By employing only SIFT features for detection, tracking and re-identification, the proposed system is most independent of the employed sensor. Unlike most other re-identification approaches [6, 3, 14], the ISM re-identification does not employ sensor specific features like color, which makes it applicable for the case of data acquired in the visible and infrared spectrum. Moreover, not using color makes this approach more robust to inter-sensor variations like different color schemes (within one spectrum).

(ii) The multi-stage approach with increasing computational cost allows for very efficient re-identification since the computational cheap first stages can be used to reduce the amount of data (candidate models from the database) that has to be considered on the last stage.

(iii) Compared to most other state-of-the-art approaches like [2, 3], this approach is applicable in real applications since it is integrated with a detection and tracking strategy. Particularly, it does not rely on manual annotation of people like [2, 3] and builds models for re-identification online without an offline training step like [1, 5].

Beyond that, this approach can be used for efficient re-identification in multi-camera network where the evolving smart cameras might build an essential part of the system. Here, person detection and tracking might be performed on the smart camera and only the tracking results are to be sent

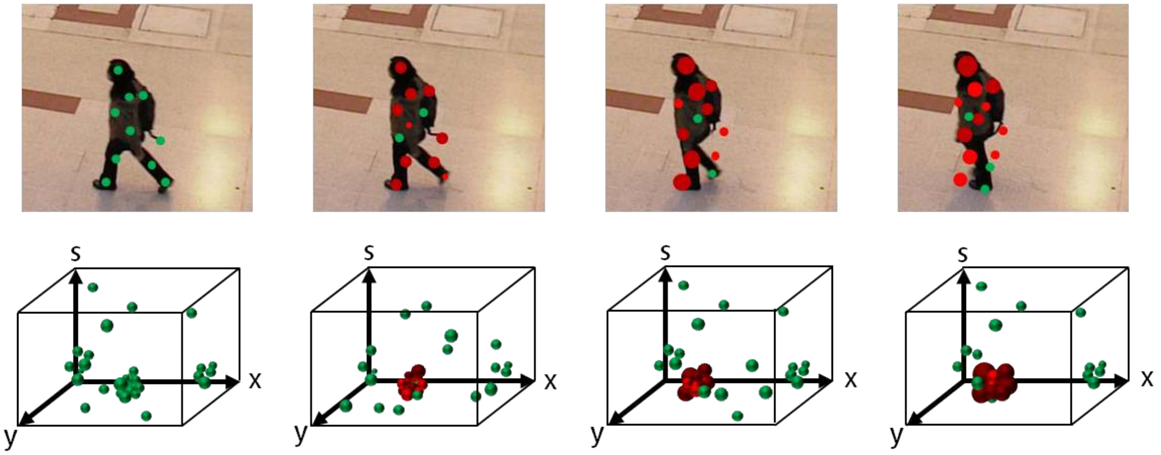

Figure 1. Short-term person model generation during tracking: Green dots visualize new features integrated into the model at this time instant while red dots indicate features which are currently perceived in the image but also have been perceived at a former time instant. Features which can not be confirmed by data (light red) for a certain duration are removed from the short-term person model.

over the network to a central computing and storage unit. In this case, not the whole video stream has to be transmitted over the network, but only the tracking results and person models which are necessary for further interpretation. The re-identification approach considered here fits perfectly into this concept because (i) The models are built online during tracking without the need for time and computational expensive offline training of classifiers that have to be outsourced to a (computationally fast) central unit and thus demand transmission of image data over the network. (ii) For each person, the whole appearance information is integrated in a single model during tracking – thus for the whole track of a person, only a single model has to be transmitted over the network (distributed to the other network nodes). (iii) The multi-staged re-identification approach allows for person representation by low dimensional models on the first two stages. Thus, only little data has to be transmitted over the network for stage 1 and 2 re-identification. Thus, these stages can be used very efficiently even when a continuous transmission of results is necessary, which might be needed for live (3D) tracking with multiple cameras with overlapping fields of view. In addition, these models demand only little storage which is relevant in cases where large public areas like airports or train stations are monitored and thus a large amount of data is to be stored.

In this paper, we use the re-identification approach formerly used for person re-identification in infrared with a single camera and show that this approach is applicable for the task of person re-identification in the visible spectrum as well. Specifically, we show that it can be successfully employed for re-identification in a multi-camera environment where strong variations in environmental conditions and camera view exist between cameras. Furthermore we compare the re-identification performance of the 3 stages and show that re-identification is possible even with low dimensional models which is particularly relevant when considering camera networks.

The paper is structured as follows. In section 2, we describe generation of person models during tracking. In section 3, we give an overview of the person re-identification approach. Section 4 presents the evaluation and section 5 concludes.

2. Tracking and model building

The tracking approach this work builds on was introduced in [7, 9]. The key idea of this tracking is that the Implicit Shape Model (ISM), which is a trainable object detection approach that builds on local features (we use SIFT throughout this paper), is extended for tracking. The ISM object detector and thus the tracking works on the basis of a codebook for a specific object category which is built in a training step based on sample images of the relevant object category. Here, SIFT features found in the training samples (SIFT features which in this case represent the category ”person”) are input to a clustering that identifies reoccurring features which are relevant for the object category. The cluster centers are employed as prototypes for the codebook which describes the object category generically. Together with these prototypes, a spatial distribution, which encodes the position of features which contributed to this prototype relatively to the object center, is stored for each prototype. This codebook is used to detect persons in input images by matching SIFT features extracted in the input image to the prototypes. The spatial distribution of matching prototypes is then employed to build a Hough voting space where each spatial distribution entry votes for a possible object center location. A mean-shift maxima search in this Hough space is conducted to detect persons. Tracking builds on this ISM detection and extends it by integrating temporal information into the Hough based object detection approach. Besides stable tracking through short-term occlusions, the essential point of this tracking is that it automatically builds SIFT feature models of the tracked persons during tracking.

Generation of these 'short-term' features models are visualized in figure 1. Circles in top-row visualize features in the person hypothesis during tracking. Bottom-row visualizes the Hough-voting space for successive frames of an image sequence. Different colors refer to the three feature types which are determined in the data association step between current and last frame: Dark red circles indicate features where a correspondence between a feature in the person model of time T-1 and a new image feature of time T could be established. Weight of these features is increased every time a valid correspondence could be established. Light red circles show person model features which have no feature correspondence in the current frame. The weight of these features is decreased in every time instant. Features the weight of which is too low, are removed from the feature model. Green features visualize image features without a correspondence with an feature in the person model of time T-1. Tracking of people is pursued by mean-shift in the Hough voting space. An important point in this tracking is, that at time T, a hypothesis specific mean-shift search is conducted for every person known at time T-1. This hypothesis specific search only includes votes for this specific hypothesis (in this case dark and light red) and votes generated from image features without a correspondence in the hypothesis feature set (green). The weight of a feature with a correspondence is increased in every time instant this specific feature has a correspondence. Thus the weight of the feature and therewith the influence in the voting process and thereby in the object detection is increased too. When no correspondence with a new image feature could be made for a certain feature in the person model, the weight of this feature is decreased. This means, when no new image evidence is available for a feature in the person model, its influence in the model is decreased and it is removed from the model when its weight is below a certain threshold. Besides the adaption of feature (and vote) weight and therewith the adaption of influence of features in the person model, another central point is that new features (green) are automatically integrated into the model. This is very important because the model has to be adaptive to appearance changes of the person that happen during tracking. In this case, appearance changes are automatically integrated into the model.

For person re-identification, it is desirable to integrate all available appearance information of a person. Since the described short-term tracking person models are volatile in such sense, that new information is integrated and obsolete information is removed from the person model continuously, for re-identification these short-term person models are extended to long-term models which integrate and keep all appearance information available for this person. The generation of these long-term models is depicted in figure 2. Here, the left hand side shows some features collected during tracking (or rather the image patches the features were calculated on). All these features are integrated into the long-term model which is shown in the middle of figure 2.

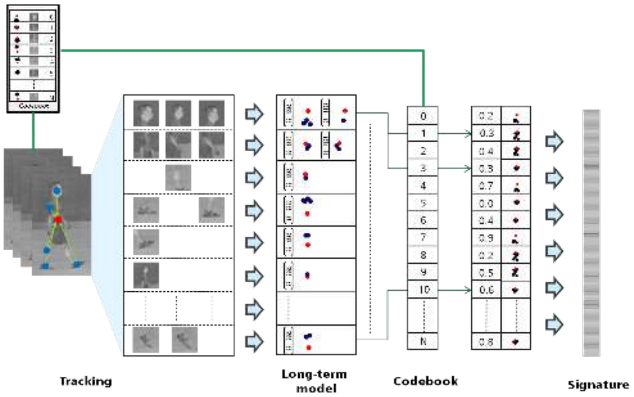

Figure 2. Long-term person model generation: SIFT features found on a person during tracking (short-term model) are transfered to the long-term model when they meet the requirements (e.g. time stability). The long-term model contains all appearance information over the whole track. To reduce amount of data that has to be stored, clustering is performed in this long-term model. In addition to the high-dimensional feature descriptors, ISM-signatures are stored for a person. As shown right hand side, these contain activation signatures and the spatial feature distributions.

In addition to the long-term model which includes all features harvested during tracking, a so called ISM-signature is stored for each person. This signature is shown on the right hand side of figure 2. This signature is build on the basis of the codebook activation of all features in the long-term model. Codebook activation in this case refers to the vector of codebook dimension N that reflects the match of a SIFT feature to the codebook prototypes. This activation is calculated during object detection by matching the feature to all codebook prototypes. This activation is now used to represent a person's appearance. For this, every feature inside the long-term model contributes to the calculation of the signature. Entry n of the N-dimensional signature is calculate as follows:

$$\Theta_n = \sum_{i=0}^{I} \theta_{i,n}. \tag{1}$$

Here, Θ_n is the n-th signature entry, I is the number of features in the long-term model and $\theta_{i,n}$ is the activation strength of the i-th feature in the long-term model of codebook entry n. This signature thereby represent the person

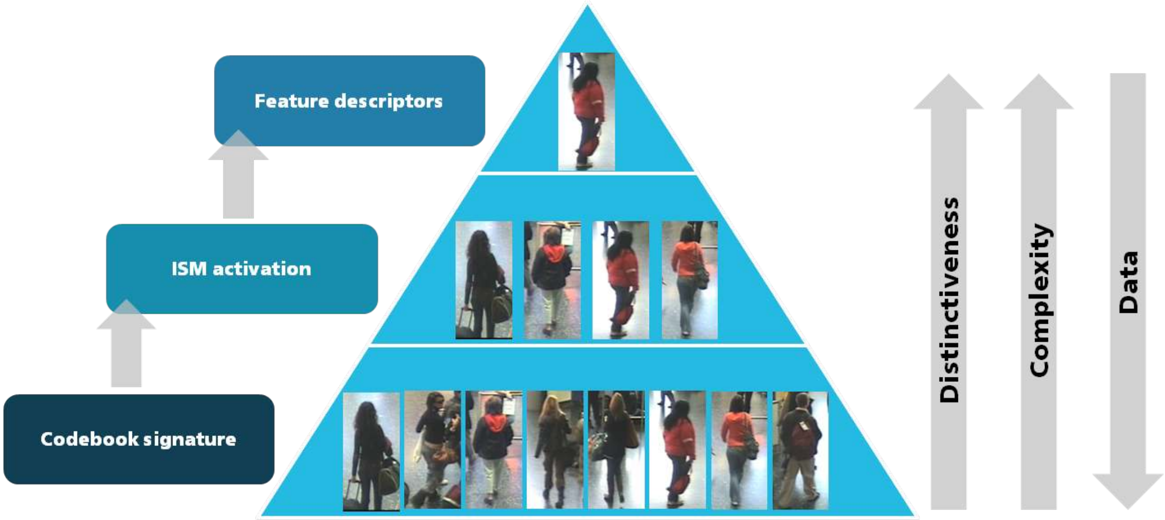

Figure 3. Overview of the 3-staged re-identification approach. In stage 1, the codebook signatures (activation signatures) are compared. Second stage uses ISM activation by adding spatial feature distribution. Third stage compares the high dimensional SIFT descriptors. By using the stages in a cascade amount of persons that have to be considered can be reduced in each stage. Thus, in the computational most expensive stage 3 which provides the highest distinctiveness, only few database models have to be considered for a query model.

appearance in form of a "bag of words" model. In addition to the overall activation strengths, for each codebook entry, the spatial feature distribution is stored. By that, not only the appearance is modeled but the position of the feature in the coordinate frame of the person too. An important point is, that both, the feature models and the ISM-signatures integrate the appearance information over the image sequence in a single model. This means, that regardless of the duration a person was tracked, only a single model which encodes the whole information has to be stored.

3. Re-identification

Person re-identification builds on the feature models and ISM-signatures acquired during tracking. These are used in a 3-staged approach that allows for efficient re-identification [8, 10].

As depicted in figure 4, in the first stage, the ISM-signature activation profiles are compared. This means that for each comparison of a query model with a database model, only two N-dimensional vectors are to be compared. N is the codebook dimension which is usually in range [100-2000]. In the second stage, the spatial distributions, which encode feature positions, are added to the comparison. This adds $N * K^2$ comparisons (where K is the mean number of entries in a spatial distribution and is around 10) and increases representation distinctiveness. In the third stage, the high-dimensional SIFT feature models are compared. This is the most expensive stage concerning computational cost but has the highest model distinctiveness too.

As the overview in figure 3 shows, the stages can be used in a cascade which allows for efficient re-identification. Every stage can reduce the number of models that has to be considered in the next stage. The stages are specifically suited for this because stages 1 and 2 provide computa-

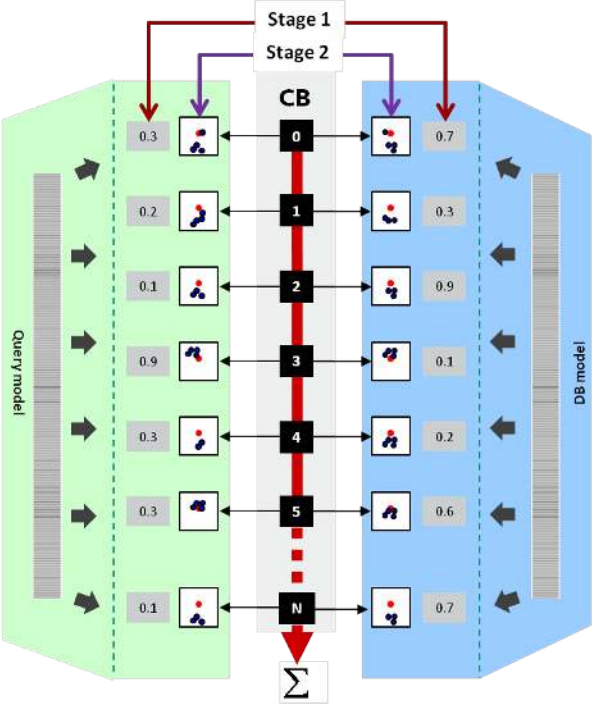

Figure 4. Person re-identification in stage 1 and 2. In stage 1, activation signatures are compared. In the second stage, spatial distributions are added to the comparison.

tional efficiency since only low dimensional data is considered here. In these stages, many of the models in a database can be removed from further consideration. Thus, computationally expensive comparison in the third stage has to be performed only for few models. Another possibility is to use stages 1 and 2 directly to perform re-identification. In this case, only low dimensional ISM-signatures have to be stored. This can be useful in cases where storage capacity or network bandwidth is limited.

4. Evaluation

We evaluate our re-identification approach in the iLids scenario [13] which is a real-world multi-camera scenario. It contains sequences recorded at an airport in official hours, which means it contains realistic surveillance data and thus all difficulties that are accompanied by this for both, tracking and re-identification. For re-identification evaluation, we use the data of two cameras with disjoint views. Sample images of the two cameras are shown in figure 5. As one can see, many challenges arise for person re-identification here, especially when seeking an evaluation under realistic conditions and thus a system that can be applied to real-world scenarios. This means, that for re-identification, we cannot assume that tracking results are flawless. In real-world applications this is only rarely the case. Specifically under the challenging conditions here, where multiple persons move through the scene and occlude each other, assuming the output of a tracker to be perfect, like many other re-identification approaches do, is unsustainable. In addition, many challenges for re-identification itself arise here: (i) people are partially occluded by luggage which affects person appearance, (ii) cameras observe the scene from different viewpoints, (iii) environmental conditions like lighting differ significantly between cameras and (iv) there are differences in the color representation between the cameras (which in fact does not affect our SIFT based approach).

Evaluation is performed on a set of 45 persons visible in both cameras. For evaluation, the database is filled with all models from one camera. All models from the other camera are used as query models. For performance measuring, we use the *Cumulative Matching Characteristic (CMC)* and *Synthetic Disambiguation Rate (SDR)* (see [4] for details). These are the de-facto standard characteristic for re-identification evaluation in challenging scenarios. CMC treats re-identification as a ranking problem, which means it accounts for the rank of the correct database model for a query model. SDR shows re-identification performance for different database sizes and thus can be used to compare re-identification approaches that have been evaluated on databases of different sizes. SDR can be computed using CMC:

$$SDR(M) = CMC\left(\frac{N}{M}\right). \quad (2)$$

Here, N is the database size that was used for CMC calculation, M the database size for SDR calculation and CMC(k) the cumulative matching characteristic for rank k.

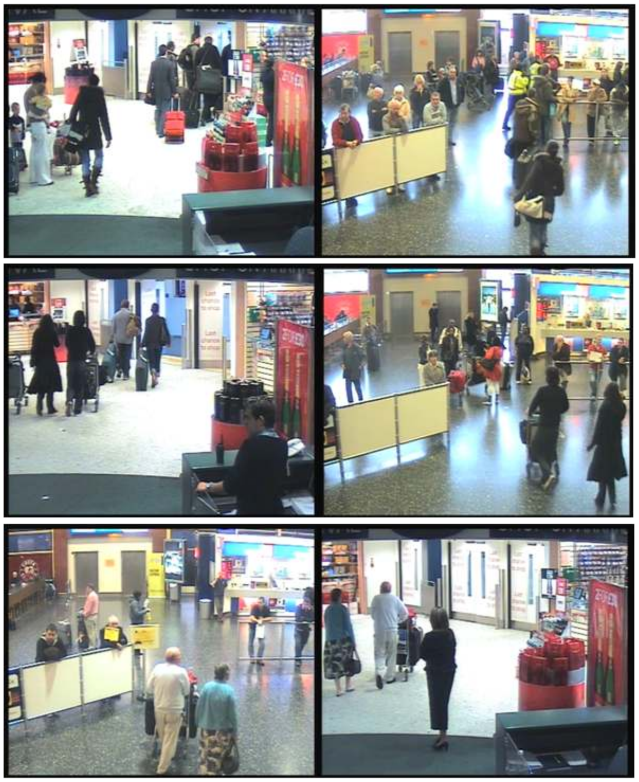

Figure 5. Sample images of cameras 1 and 3 of the iLids scenario. As one can see, strong differences in environmental condition, e.g. changes in lighting, are present between the cameras. Other challenges, e.g. occlusion of persons by other persons and luggage increase severity for both, tracking and re-identification.

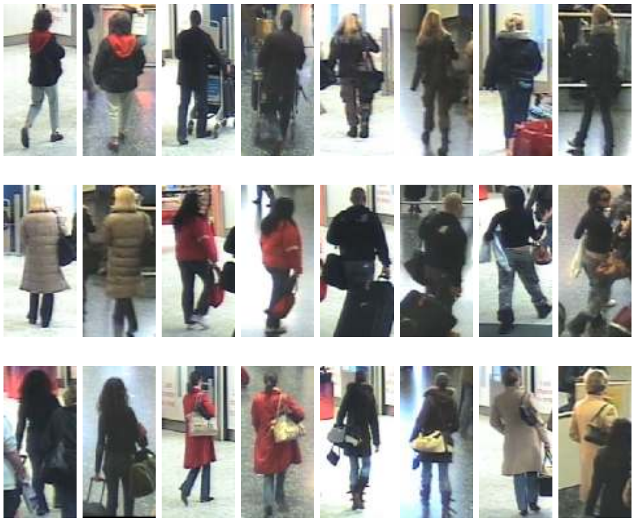

Figure 6. Sample persons of the iLids scenario.

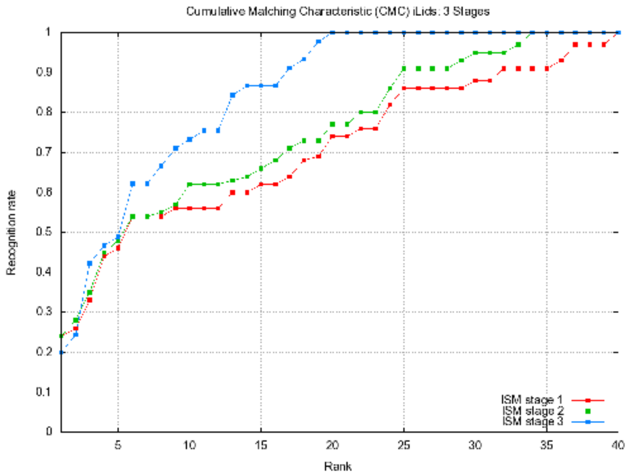

Figure 7. Cumulative matching characteristic for the three stages of ISM re-identification in the iLids scenario.

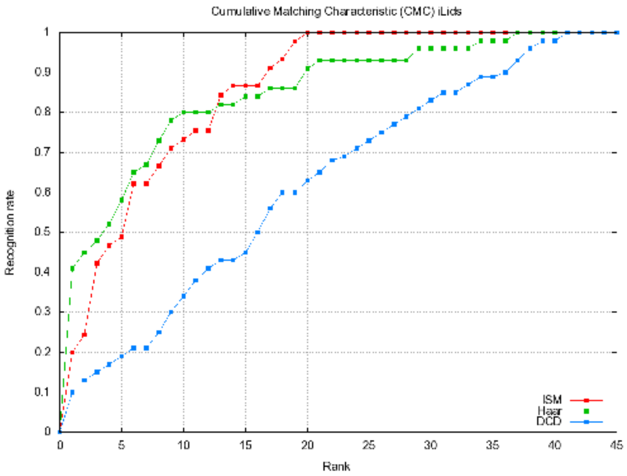

Figure 8. Cumulative matching characteristic of ISM re-identification and two other appearance based re-identification approaches in the iLids scenario.

Performance in the iLids scenario is shown in figure 7 for the three ISM re-identification stages. As we see here, performance on all three stages is very similar up to a rank of 5. This shows that even the low dimensional description in stages 1 and 2 can be employed to perform reasonable person re-identification. This is important for cases that disallow transmission of high dimensional models, e.g. in camera networks where bandwidth is limited. Starting from rank 6, stage 3 performance is superior to stages 1 and 2. This is due to the higher model distinctiveness of stage 1 that permits spotting of detailed varieties. The difference between stage 1 and 2 is only marginal. This means that using the spatial distribution in stage 2 only increases distinctiveness slightly. This is due to the view-differences between cameras which disallows demand of high similarity between spatial distributions.

Comparison of ISM re-identification with two other appearance based re-identification approaches from [1], which have been evaluated on data from the same dataset (44 persons) is shown in figure 8. 'Haar' uses Haar-like features for person description, 'DCD' uses dominant color descriptors. Both approaches use an adaboost trained classifier for re-identification. The reason why we choose these approaches for comparison is that both do not use artificial data for evaluation, i.e. manually cut images of persons as basis for re-identification, but use real tracking results as we do. Thus a fair comparison is possible.

As we see from figure 8, performance of DCD is far behind performance of the other approaches. This is due to the strong environment and camera type induced color differences between the cameras, which are not (see [1]) resolvable by color calibration between cameras. Performance of ISM re-identification is comparable with Haar performance. In the first half of the graph, Haar performs slightly better than ISM. We think that this is due to the offline training step that is used in Haar re-identification and which allows for a more distinct model compared to ISM re-identification where models are built online. In the second half of the graph, ISM performance is slightly above Haar performance.

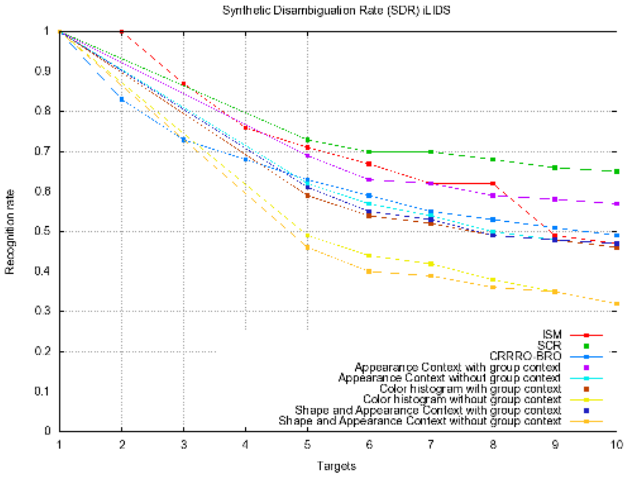

Figure 9. Synthetic disambiguation rates of ISM re-identification and other appearance based re-identification approaches in the iLids scenario.

To allow for comparison with approaches that use a different database size, we use SDR which shows re-identification performance as a function of database size. SDR is a good indicator for suitability of re-identification for multi-camera tracking because in multi-camera tracking, information about spatio-temporal dependencies between cameras can be used to reduce the number of persons that have to be considered for appearance based re-identification.

For the comparison shown in figure 9 we picked approaches that have been evaluated on the same iLids dataset but used different database size. In detail, following approaches are shown: *Spatial covariance Regions (SCR)* [2], *Center Rectangular Ring Ratio-Occurrence Descriptor-Block based Ratio-Occurrence (CRRRO-BRO)*, *Appearance Context*, *Color histogram* and *Shape and Appearance Context* from [15]. Last three approaches have been evaluated with and without group context. Note that other approaches shown here do not build on a real tracking but use manual annotation of persons for re-identification. Since these approaches do not have to deal with real difficulties that arise from occlusion during tracking etc., comparison with these approaches is somehow biased. Despite that, ISM re-identification shows at least comparable performance. With few persons, ISM re-identification outperforms all approaches except SCR. Starting from 9 persons, three approaches outperform ISM re-identification. In our opinion, this is due to the offline training step the other approaches perform and which increases distinctiveness. We do not perform such an offline training step because it dismisses system applicability in cases where online processing is necessary (e.g. in multi-camera-tracking).

5. Conclusion

In this paper, we approached the task of person re-identification for multi-camera networks. For that we outlined an integrated approach for person tracking and re-identification which is based on the Implicit Shape Model and SIFT features only and thus generically applicable. Evaluation in a challenging real-world scenario has shown that ISM re-identification which (i) is solely based on SIFT and thus applicable independently of sensor modality, (ii) builds on a real tracking and thus is applicable in real-world scenarios and (iii) performs integration of temporal information in a single model and thus allows for efficient storage and matching of person models, provides comparable performance as specialized re-identification approaches with a narrow applicability range.

References

[1] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using haar-based and dcd-based signature. In *Proc. Advanced Visual and Signal based Surveillance*, pages 1528–1535, 2010. 57, 62

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. Advanced Video and Signal based Surveillance*, pages 435–440, 2010. 57, 63

[3] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, 2006. 57

[4] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. International Workshop on Performance Evaluation for Tracking and Surveillance*, pages 1–8, 2007. 61

[5] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. European Conference on Computer Vision*, pages 262–275, 2008. 57

[6] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. Computer Vision and Pattern Recognition*, pages 26–33, 2005. 57

[7] K. Jüngling and M. Arens. Detection and tracking of objects with direct integration of perception and expectation. In *Proc. International Conference on Computer Vision (ICCV Workshops)*, pages 1129–1136, 2009. 58

[8] K. Jüngling and M. Arens. Local feature based person reidentification in infrared image sequences. In *Proc. Conference on Advanced Video and Signal based Surveillance*, pages 448–454, 2010. 57, 60

[9] K. Jüngling and M. Arens. Pedestrian tracking in infrared from moving vehicles. In *Intelligent Vehicles Symposium*, pages 470–477, 2010. 58

[10] K. Jüngling and M. Arens. A multi-staged system for efficient visual person reidentification. In *Proc. of the Conference on Machine Vision Applications*, pages 1–8, 2011. 57, 60

[11] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008. 57

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 57

[13] U. H. Office. i-lids multiple camera tracking scenario definition, 2008. 61

[14] D. Truong Cong, L. Khoudour, C. Achard, and L. Douadi. People detection and re-identification in complex environments. *IEICE Transactions on Information and Systems*, 93:1761–1772, 2010. 57

[15] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proc. British Machine Vision Conference*, pages 1–8, 2009. 63