A multi-staged system for efficient visual person reidentification

Kai Jüngling
Fraunhofer IOSB
Ettlingen, Germany
kai.juengling@iosb.fraunhofer.de

Michael Arens
Fraunhofer IOSB
Ettlingen, Germany
michael.arens@iosb.fraunhofer.de

Abstract

An important field in today's computer vision is person centric video analysis. The basis of this person centric analysis is the detection and tracking of people in video data. In many cases it is not sufficient to track people when they continuously appear in the camera's field of view, but to also reacquire a track after a person has left a field of view and reenters it. In this paper, we introduce a technique that conducts this person reidentification based on SIFT features only. This approach fits into an Implicit Shape Model (ISM) based person tracking approach by employing the SIFT features collected during tracking for reidentification. The ISM characteristics of a person are used to perform reidentification in an efficient 3-staged approach which combines computation efficiency with high distinctiveness. The evaluation is performed in an open-set classification approach on a public dataset of 60 persons which was acquired with a thermal camera. Despite the challenges of person reidentification in thermal imagery, the approach shows nearly perfect performance and outperforms other reidentification approaches on this dataset.

1 Introduction

Object, and more specific person tracking is an indispensable part of many of today's computer vision applications. In many cases, specifically in areas like visual surveillance, tracking of a person while it continuously appears in the camera's field of view is not sufficient to build a comprehensive image that allows for high level video analysis. To build a spatially and temporally comprehensive representation, a person has to be reidentified when reentering the camera's field of view.

In this paper, we tackle this task of person reidentification and propose an approach which is integrated in a detection and tracking framework and by that seeks to be applicable in real-world applications. For that, we build on the Implicit Shape Model (ISM) based person detection and tracking system proposed in [7, 9]. The main idea of this paper is to closely integrate tracking and reidentification by employing the SIFT features which are used for and collected during tracking for reidentification too. Not only the SIFT features are used for reidentification, but the Tracking-ISM characteristics are used to reidentify a person in an efficient 3-staged approach. This approach consists of levels of increasing distinctiveness and complexity. The arrangement of the 3 stages in a classification cascade gains strong distinctiveness in reidentification while being computational efficient. The overall approach has several advantages over existing reidentification approaches:

(i) By employing only SIFT features for detection, tracking and reidentification, the proposed system is *most independent of the employed sensor*. Unlike most other reidentification approaches [6, 4, 11], the ISM-reidentification does not employ sensor specific features like color, which makes it applicable for the case of data acquired in the visible and infrared spectrum.

(ii) The multi-stage approach with increasing computational cost allows for very efficient reidentification 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 [3, 4], this approach is applicable in real applications since it is integrated with a detection and tracking strategy. Specifically, it does not rely on manual annotation of people like [3, 4] and builds models for reidentification online without an offline training step like [2, 5].

With this approach, we follow the idea of Jüngling and Arens [8], but extend the approach to increase reidentification distinctiveness and to overcome inconsistencies in the model building approach. With these extensions, we are able to distinguish more people and increase reidentification performance. This is documented by an evaluation of the reidentification on a dataset of 60 persons, which is an increase of more than 100% (35 persons) compared to [8]. In contrast to most other approaches, we here perform an open-set classification approach which is the most challenging task for reidentification but conforms with the needs in real applications. Comparison of the results with those of other reidentification approaches on the same dataset show that the ISM-reidentification clearly outperforms those other approaches on this dataset. This paper is outlined as follows. Section 2 introduces the 3-staged reidentification approach. Results of the experiments can be found in section 3. The paper is concluded is section 4.

2 Person reidentification

We adopt the tracking and detection strategy introduced in [7, 9]. The tracking described there detects and tracks persons by matching SIFT features extracted from an input image with an appearance codebook. The important aspect for reidentification is, that, during tracking, short-term feature models of the tracked persons are built. For reidentification, these are extended to long-term-models that comprise the whole appearance information of a person. As visualized in figure 1, these long-term identity models each contain a number of feature clusters which are built during tracking. In addition to a SIFT descriptor that models the appearance of the cluster, the ISM charac-

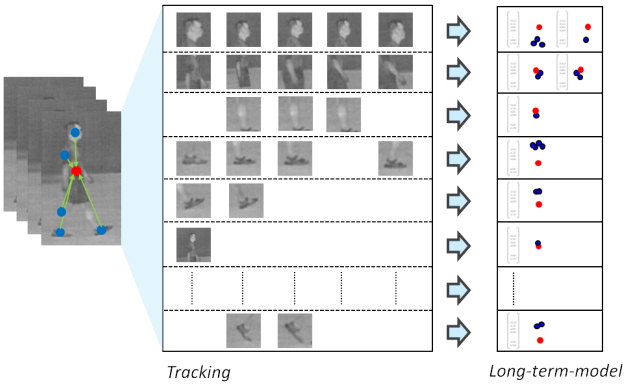

Figure 1. Identity model generation during tracking.

teristic is stored for each cluster. This characteristic comprises the spatial distribution of features that contributed to the cluster in terms of object center offsets and the codebook activation vector of the contributing features.

Using that information, person models are compared for person reidentification in a 3-staged approach. Distinctiveness is increased in each subsequent stage by including additional information.

Stage 1 uses *codebook activation signatures* which are built during tracking for person reidentification. Here, low dimensional (codebook dimension is 216 in the context of this paper) signatures are to be compared for reidentification. This stage has very low computational cost but is rather limited regarding distinctiveness. Distinctiveness is increased in *stage 2*, where *spatial feature distributions* are included in reidentification and thereby, the shape of a person is modeled. Here, computational cost is only increased slightly. In *stage 3*, *SIFT feature descriptors* are matched. This stage includes the highest distinctiveness but the biggest computational complexity too. Since the computational cheap stages 1 and 2 can be used to discard a lot of database models (in case this approach is used to compare a query model to a database of person models), stage 3 only has to be carried out for a fraction of all models. Thus the higher complexity on this stage is acceptable.

2.1 Stage 1: Codebook signature

The codebook signature of a person is built by combining codebook activations of long-term model feature clusters. The codebook activations of the feature clusters are gained during person detection (see [7, 9] for details) and describe a feature in terms of "visual words" based on SIFT descriptors. The n-th signature entry Θ_n is built by summing the activations strengths of all I clusters θ_i:

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

Using these signatures, the match of two person models ζ and η is the sum of codebook activation differences:

$$\chi_1(\zeta,\eta) = \frac{1}{N}\sum_{n=0}^{N}\left(X - |\frac{|\zeta_n|}{\zeta_T} - \frac{|\eta_n|}{\eta_T}|\right). \tag{2}$$

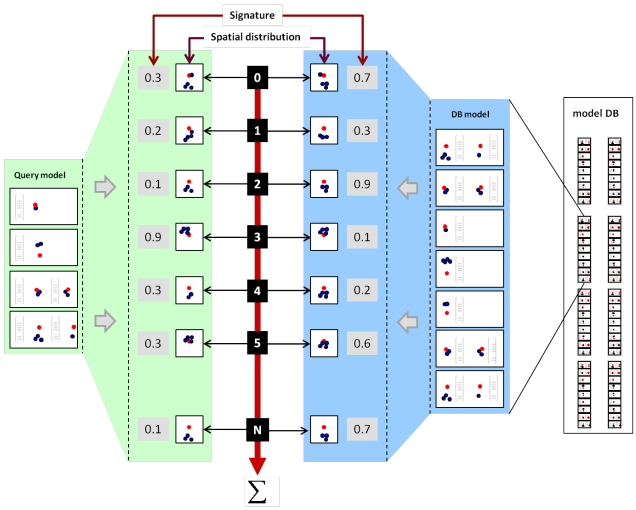

Figure 2. Model matching in stage 2: codebook signatures and spatial feature distribution are matched.

The differences are normalized with the tracking durations ζ_T and η_T respectively. X is a constant that is used to convert the distance into match. The choice of this constant is uncritical.

2.2 Stage 2: ISM activation

The second stage uses the whole ISM-characteristic for model matching. This means, that in addition to the codebook activation of the first stage, the spatial feature distribution is used in matching. By that, feature model distinctiveness is increased strongly while matching complexity is only increased slightly – each codebook entry has about 10 entries in the spatial feature distribution which results in 10^2 comparisons of 2D positions.

The match of two models ζ and η here is determined by:

$$\chi_2(\zeta,\eta) = \frac{1}{N}\sum_{n=0}^{N}\left[\left(X - |\frac{|\zeta_n|}{\zeta_T} - \frac{|\eta_n|}{\eta_T}|\right)\cdot \beta_S(\zeta_n,\eta_n)\right], \tag{3}$$

with $\beta_S(\zeta_n,\eta_n)$ being the matching condition for the spatial distributions:

$$\beta_S(\zeta_n,\eta_n) = \begin{cases} 1, & \text{if } \min_{i,k}(dist_{eukl}(\zeta_i,\eta_k)) < \delta_S^{MAX} \\ 0, & \text{else} \end{cases}. \tag{4}$$

δ_S^{MAX} defines the upper boundary for spatial distance. At least one pair in the feature distributions must have an euclidean distance below that upper boundary. Otherwise, $\beta_S(\zeta_n,\eta_n)$ in equation 3 and thus the activation match for this codebook entry is 0. Note that the activation differences in equation 3 have already been computed for all database models in stage 1. Thus, only the comparison of spatial distributions has to be calculated here.

2.3 Stage 3: SIFT descriptor

In stage 3, the SIFT cluster descriptors of the person models are compared. Thus, this stage has the

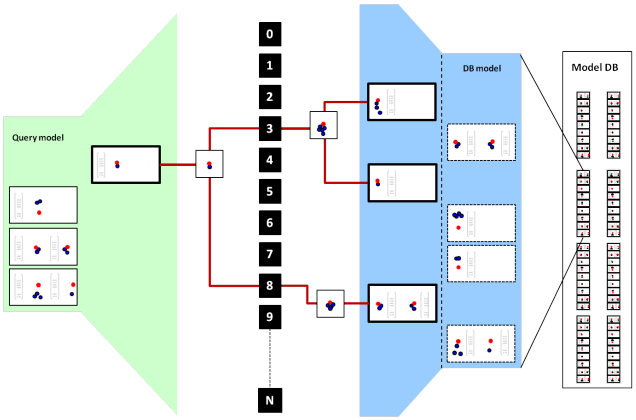

Figure 3. Model matching on stage 3: SIFT descriptors are matched.

highest person description distinctiveness. The higher distinctiveness comes with increasing computation demand since sets of 128-dimensional descriptors have to be compared. The first and second stage can be used to filter out dissimilar database models and by that reduce the computational cost on the third stage by reducing the number of database models that are to be matched here. In addition to that, computational complexity in stage 3 can be reduced by employing the codebook to index features in model matching. This is shown in figure 3. For every model cluster of the query model, the best matching cluster of the database model is picked. The codebook activation signatures of the query model feature cluster is used to choose the clusters from the database models that are to be compared. Thus, not every cluster descriptor from the database model has to be matched with a query model cluster descriptor, but only those that comply in codebook activation. For those clusters that have compliance in codebook signature, the spatial distributions are matched. Again, compliance is demanded here and clusters without a match are discarded. By that, the number of cluster the SIFT descriptors of which have to be matched can be reduced significantly.

Formally, the match $\chi_3(\zeta, \eta)$ of a query model ζ with K model clusters and a database model η is defined by:

$$\chi_3(\zeta, \eta) = \frac{\sum_{k=0}^{K}(\beta_{DS}(\zeta_k, \eta))}{\zeta_T + \eta_T}, \tag{5}$$

where ζ_T und η_T are the track durations of the query and database model respectively. The similarity $\beta_{DS}(\zeta_k, \eta)$ of model cluster ζ_k and a database model η is the minimal distance of this cluster and the database model feature clusters. Here, the match between clusters is composed by the three factors ν_{AC}, ν_S and δ_D. ν_{AC} and ν_S implement the gate functions for codebook signature and spatial distribution match respectively:

$$\nu_S(\upsilon, \psi) = \begin{cases} 1, & \text{if } \min_{i,k}(dist_{eukl}(\upsilon_i^S, \psi_k^S)) < \delta_S^{MAX} \\ \infty, & \text{else} \end{cases}, \tag{6}$$

$$\nu_{AC}(\upsilon, \psi) = \begin{cases} 1, & \text{if } \exists\, \upsilon_i^{AC} \in \upsilon_I^{AC}, \upsilon_i^{AC} = \psi_j^{Ac} \\ \infty, & \text{else} \end{cases}. \tag{7}$$

Thus, in practice, the descriptor distance δ_D, which is the squared euclidean descriptor distance, has to be computed only if both preceding factors evolve to 1. This additionally reduces the computational cost since δ_D is the computationally most expensive factor.

Figure 4. Sample persons of the CASIA C test set.

3 Evaluation

Evaluation is carried out using the same performance measures used in [8]: *False Rejection Rate (FRR)* is the ratio of query persons that are rejected by the system but in fact are in the database and models in the database. *False Acceptance Rate (FAR)* is the ratio of query person that are not in the database but are classified as a certain person in the database and models in the database. *Misclassification Rate (MCR)* is the ratio of query persons that are classified as the wrong database person and models in the database. The *Correct Classification Rate (CCR)* is the ratio of correct classifications and models in the database and by that joins MCR and FRR ($CCR = 1.0 - MCR - FRR$).

3.1 Experiments

Experiments are carried out on sequences of 60 persons of the CASIA C dataset [1]. This dataset was acquired by a thermal sensor with a resolution of 320x240 at 25 fps. As the sample images in figure 4 show, this dataset is very challenging because neither color nor rich texture are available to distinguish persons. For reidentification evaluation, we build a database of 50 persons by generating identity models of these during tracking. A second sequence of each person serves as test sequence. 10 additional sequences of persons which are not in the database serve as impostors for the open-set classification. In this open-set classification, the system has to decide, whether a person has been seen before and if so, which one. This is the most challenging task for person reidentification but complies with requirements in real-world applications.

Since the discriminative power of stage 1 and 2 is limited, in a real system, these stages are to be used only for filtering of dissimilar models. Classification should be performed in stage 3 since this provides highest distinctiveness. In our experiments, we perform reidentification in stages 1 and 2 too to assess the classification ability of these stages. Since these stages do not provide enough discriminative power to perform an open-set classification, a best-match classification without impostor samples is performed here. Results are reported for evaluation on single frame and on sequence basis. For sequence classification, additional temporal consistency demands are included. For stage 1 and 2, the sequence is classified as the database model with the highest single frame match count. In stage 3,

Table 1. CASIA C ISM-reidentification results.

Stage	Type	FAR	FRR	MCR	CCR
1	Image	-	-	26.9	73.1
	Sequence	-	-	14.0	**86.0**
2	Image	-	-	24.0	76.0
	Sequence	-	-	8.0	**92.0**
3	Image	12.2	4.0	0.8	95.2
	Sequence	0	0	0	**100.0**

where open-set classification is performed, a temporal consistency of 50% is demanded. This means, a classification decision is made only if a query track is classified as a certain database person for a coherent interval of 50% the track duration. Otherwise, the query track is rejected as unknown person. The classification thereby is based on the ratio of best and second best database match. If this ratio exceeds 1.4 for a frame, a classification decision is made, otherwise this frame is counted as unknown person.

Table 1 shows the performance of *stage 1* reidentification. One can see, that the performance on this stage is very good with a CCR of 73% for single frame and 86% for sequence classification. Although this is not an open set classification, this is a very good performance since this stage has very low computational costs – only a single vector of size 216 has to be compared for every database model. Another important fact is, that the correct person is always within the top 5. This means, that this stage is perfectly suited to serve as a filter. In this case, 90% of the models can be filtered and by that reduce the computational cost in subsequent stages.

In *stage 2*, CCR increases by 3% (single frame) and 6% (sequence) compared to stage 1. This is due to the higher model distinctiveness that is gained by inclusion of the spatial distribution.

In *stage 3*, an open-set classification is performed. Under these more difficult circumstances, single frame CCR is 95.2% at a FAR of 12.2%. When considering the whole sequence, the CCR increases to a perfect classification rate of 100% with a FAR of 0%.

Since no other appearance based approaches exist which tackle the more challenging case of thermal data, we compare these results to the gait-recognition-based person reidentification of [10] which was evaluated on the same dataset. The CCR of those approaches, the head torso image (HTI) and the gait energy image (GEI) are shown in table 2. Here, different reidentification rates are reported for different choices of training and test data. A: walking speed normal-normal, B: no backpack-backpack, C: walking speed normal-slow and D: walking speed normal-quick. In our experiments, we found that our approach is independent of these issues and the classification rates are not affected either positively or negatively by the choice of training and test data. As one can see, our reidentification approach outperforms their approach by far in stage 3. It is worth noting, that our experiments were more challenging due to the open set classification we performed. Even the computational cheap stages 1 and 2 with 86% and 92% respectively, outperform Tests B, C, and partially D of the gait-recognition approaches.

Table 2. CASIA C gait-recognition CCR rates.

	A	B	C	D
GEI	96%	60%	74%	83%
HTI	94%	51%	85%	88%

4 Conclusion

In this article, we introduced an efficient multi-stage approach for SIFT-based person reidentification. The evaluation in image sequences acquired by a thermal camera show the good performance of our approach under the difficult circumstances for person reidentification in thermal data. Comparison to other reidentification approaches show that our approach clearly outperforms these approaches even in the more challenging task of an open-set classification.

References

[1] Casia gait database, http://www.sinobiometrics.com. obtained from http://www.cbsr.ia.ac.cn/english/gait

[2] 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.

[3] 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.

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

[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.

[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.

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

[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.

[9] K. Jüngling and M. Arens. *Local Feature based Person Detection and Tracking Beyond the Visible Spectrum.* Springer, To appear 02/2011.

[10] D. Tan, K. Huang, S. Yu, and T. Tan. Efficient night gait recognition based on template matching. In *International Conference on Pattern Recognition*, volume 3, pages 1000–1003, 2006.

[11] D.-N. 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.