2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

Local Feature Based Person Reidentification in Infrared Image Sequences

Kai Jüngling, Michael Arens
Fraunhofer IOSB
Gutleuthausstrasse 1, 76275 Ettlingen, Germany
{kai.juengling, michael.arens}@iosb.fraunhofer.de

Abstract

In this paper, we address the task of appearance based person reidentification in infrared image sequences. While common approaches for appearance based person reidentification in the visible spectrum acquire color histograms of a person, this technique is not applicable in infrared for obvious reasons. To tackle the more difficult problem of person reidentification in infrared, we introduce an approach that relies on local image features only and thus is completely independent of sensor specific features which might be available only in the visible spectrum. Our approach fits into an Implicit Shape Model (ISM) based person detection and tracking strategy described in previous work. Local features collected during tracking are employed for person reidentification while the generalizing appearance codebook used for person detection serves as structuring element to generate person signatures. By this, we gain an integrated approach that allows for fast online model generation, a compact representation, and fast model matching. Since the model allows for a joined representation of appearance and spatial information, no complex representation models like graph structures are needed. We evaluate our person reidentification approach on a subset of the CASIA infrared dataset.

1. Introduction

A common task in visual surveillance is the tracking of persons. In many cases, it is not sufficient to track a person when it continuously appears in the camera's field of view, but to also determine if a person that enters the cameras field of view has been seen before. In typical surveillance situations, this reidentification of persons cannot be conducted by using common biometric approaches like face recognition, because the persons faces may not always be visible and the camera resolution may not be sufficient to allow for face recognition at all. In these cases, person reidentification can be performed just on global appearance information of the person assuming that a person wears the same clothes in the time period that is relevant for reidentification. In this paper, we focus on reidentification situations where this assumption holds. Since the ability for surveillance at night becomes more and more important, the use of infrared cameras for surveillance applications increases, too. This is only consequent because typical surveillance scenarios mainly involve the surveillance of people which makes thermal sensors well suited. For that reason, surveillance tasks like person reidentification have to be tackled in infrared data too. In this paper, we face the task of person reidentification in infrared and present an integrated approach, that uses local image features for person detection, tracking and reidentification.

While some approaches focus on person tracking in infrared [9, 4, 13, 18], only little research tackles the task of appearance based person reidentification in infrared. This is most likely due to the inherent difficulties for person reidentification that exists here. Most person reidentification approaches for the visible spectrum focus on using color, especially color histograms, for object reidentification (e.g. [15]). Here, a lot of effort has been put into building color models that can be used to track people in camera networks [16, 8, 7]. Besides the drawback of relying on an object segmentation (which is only obtainable when making restriction on the application scenario), these approaches are obviously not applicable in infrared data. Some of the approaches proposed for the visible spectrum rely on local image features and thus might be applicable for infrared data too: Hamdoun *et al*. [6] proposed a person reidentification approach based on SURF (Speeded Up Robust Features [3]) like features. Here, person reidentification is carried out by a KD-tree. This allows for fast matching of person models and thus efficient database queries. A drawback that we see in this approach is that no spatial information of features, and thus no dependencies between features are used. In addition, this approach does not seem to be integrated in a tracking framework as they evaluate their system with (as it seems) hand selected frames. In their approach, there is no means to create a distinct model since the feature selection seems somehow random. Gheissari *et al*. [5] propose

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.75

a person reidentification approach which uses a combination of salient edgel and color histograms. Arth *et al*. [2] introduce a related reidentification approach for cars. Here, visual words are used to built an object fingerprint which is used for comparison with other instances of the same object class (here cars).

In this work, we propose a local feature based person reidentification approach that is related to the approach of Arth *et al*. [2] in such a way that we use visual words for person reidentification. Since people are articulated objects and usually have only little structural differences (unlike different types of cars) which can be spotted using bag of features like methods, we introduce an extended model that suits the needs in person reidentification. In addition, our reidentification approach is integrated with a local features based person detection and tracking approach and thus completely self contained and most independent of specific application scenarios (e.g. the overall approach is applicable for moving cameras, too). By using only SIFT [14] for all three tasks, the approach is most independent of sensor specifics too - the overall approach can be applied to data in the visible spectrum without modifications. In contrast to approaches like [17], where an explicit feature based graph representation is build for person tracking based on local features, our approach builds on an Implicit Shape Model (ISM) based person detection and tracking strategy. The codebook used for person detection can be used for indexing the local features which are collected during tracking for person reidentification. By using this general appearance codebook as a basis structure for person instance models, we gain an integrated approach that allows for fast online model generation, a compact representation and fast model matching. Since it allows for a joined representation of appearance and spatial information, it makes more complex feature representations like graph structures [17] unnecessary in our context. Another major advantage of our approach is that a reidentification decision can be made at every point during tracking and no artificial frame or sequence selection has to be applied like in other reidentification approaches. This makes the integrated detection, tracking and reidentification approach well suited for real world surveillance tasks.

In what follows, section 2 gives an overview of the detection and tracking approach we build on. Section 3 introduces our person reidentification strategy, which is evaluated in section 4 for infrared image sequences. Section 5 concludes this paper.

2. Person detection and tracking

2.1. Detection

We build our work on the infrared pedestrian detector described in [11] (see [12] too). A brief overview of this detection approach is given in Fig. 1. In a *training stage*, SIFT features are extracted from training samples. After a clustering stage where feature prototypes are built, an Implicit Shape Model (ISM) records the spatial occurrence of features in terms of object center offsets. This non-parametric feature distribution together with the appearance prototypes build the codebook for the trained object class. To *detect* objects in input images, the codebook prototypes are matched with SIFT features extracted from the input image. Matching features cast votes for object center locations (determined by the training ISM) in a three dimensional Hough voting space comprising two dimensions (x,y) for image location and one (s) for scale. To find object hypotheses, a maxima search is conducted in this voting space. For fast initialization of the search, initial maxima are defined by maxima in a grid partitioning and afterwards refined by mean shift. The most important part for the remainders is that object detection provides us with a set of object hypotheses which include disjoint image feature sets which lead to the hypotheses. Since all features in a hypothesis passed codebook matching to be included in a hypothesis, they are annotated with a codebook entry (index) and an object center offset. This is important for person reidentification because this information is used to build person instance models.

2.2. Tracking

The tracking approach [10] used in this context builds on the detection approach described in section 2.1 and conducts tracking based solely on local SIFT features. The principle approach works by a propagation of hypotheses on the feature level from one point in time to the next. By that, tracking – especially identity preservation – is automatically pursued and integrated into the object detection approach by fusing expectations and new data.

Most important for person reidentification is, that tracking provides person identities for temporally connected appearances of a person in the scene. Additionally, it provides SIFT features which were collected during tracking (and by that tracking was conducted). As was shown in [10], this tracking approach is perfectly suited to track person through short term occlusions. Even more important for person reidentification, tracking automatically builds and updates person models during tracking by feature propagation. These models are volatile in such way, that they are adapted continuously to integrate appearance changes of the modeled person. This means, that new features are integrated into the model and other features are removed from the model to provide the best estimate of current person state. In contrast to this short term tracking model, that integrates only the recent history, for person reidentification, a long term model that integrates the whole appearance history of a person is necessary. This model must integrate as much appearance

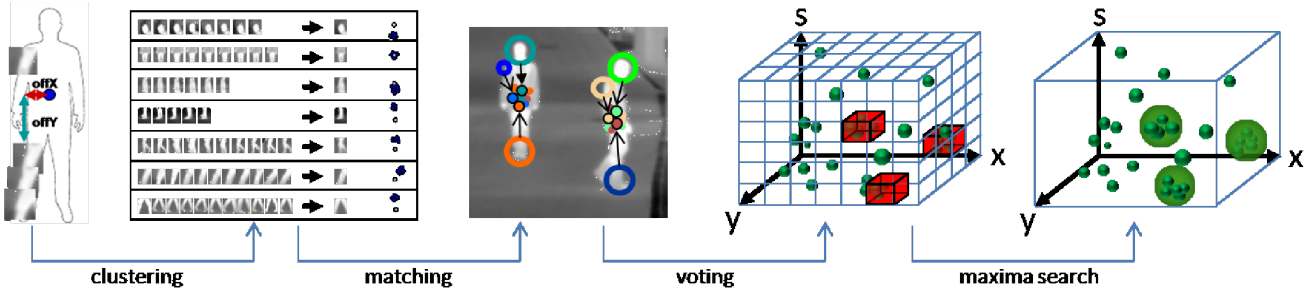

Figure 1. Overview of the object detection approach. Local features found on training samples are clustered in descriptor space to build a general person codebook. Cluster centers are prototypes for descriptors representation, object center offsets build the spatial distribution. To detect persons in an input image, local features extracted from the input image are matched with the codebook prototypes. Codebook offsets cast votes for object center locations in a 3D voting space. Maxima found in this voting space by mean shift define object hypotheses.

information as possible, since, due to viewpoint and articulation changes, the appearance of a person changes over time. To be able to reidentify a person, this versatile information has to be integrated into a single model. This is challenging due to the information amount that is collected during tracking. The goal is to store this information in an efficient representation form without loosing too much information or generalizing to much. Additionally, this representation form should be capable of serving as a basis for matching the model with other models.

3. Person reidentification

The main idea for person reidentification is to use the codebook structure as a basis to build and match person models. For that, local features which are collected during tracking are indexed using the codebook entries. This indexing serves two purposes. First, the codebook indexes serve as structural component for the model in such a way that features which have the same codebook index must have a similar visual appearance (since they have been activated by the same codebook entry in detection, the similarity of visual appearance is thus defined by the matching radius in object detection. See [11] for details.). Second, the codebook indexing of models can be used for efficient matching of feature models.

3.1. Identity model generation

During tracking, we collect features that are found in a specific person hypothesis. These features are then integrated into our person model for reidentification. For a time T we have a set of currently perceived (image) features Φ_T^ζ of a hypothesis ζ. Since all these features passed the person detection to be included in the hypothesis feature set, the involvement of every image feature resulted from a match with a specific codebook prototype. Since we only use a single vote of each image feature (see [11]), the connection between an image feature that was involved in a person hypothesis and codebook prototype is decisive. This decisive connection between image feature and codebook entry is used to build a model based on codebook indexing. For that, all features in the hypothesis feature set are assigned to the according codebook entry at every tracking step. Although the codebook can serve as structure to organize features, it does not reduce the amount of data that is collected during tracking in form of local feature descriptors. To generate a compact model representation, we build descriptor clusters in each model entry during tracking. For that, every new image feature that is added to the model is matched with all existing feature descriptor clusters in the according model entry. If the similarity to one of the clusters (represented by the cluster mean) is high enough, the feature is added to this cluster and the cluster prototype is updated (weighted mean) with the new feature, otherwise a new cluster is generated. This approach is visualized in Figure 2. The left side shows an excerpt of the person codebook. The middle part visualizes tracking and features (visualized by image patches) which are collected during tracking and their codebook entry affiliation. Note that this affiliation is a result of person detection and no additional matching step is necessary here. On the right side, we see a visualization of the model which was built in tracking after these 5 steps (in fact the model is updated continuously in every tracking iteration). This model can afterwards be used for person reidentification by matching it with other models. As we see, the number of clusters depends on the visual similarity of feature descriptors. In addition to the cluster prototypes and the number of cluster entries, the object center offsets are stored to allow for spatial consistency checking in person reidentification.

3.2. Model matching

By structuring person models with the general appearance codebook, we have a compact representation structure which can be used for fast matching as well. Using the

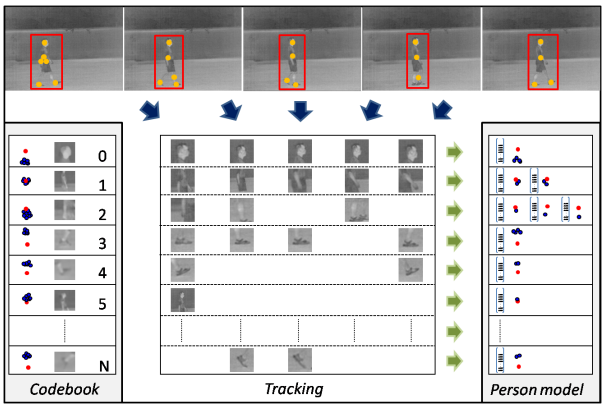

Figure 2. Person model generation during single camera tracking. SIFT features collected during training are integrated into the person instance model indexed by the codebook entries which activated the feature during detection. A model entry specific online descriptor clustering generates a compact model representation which, together with the spatial distribution of the features, allows for fast matching of feature models.

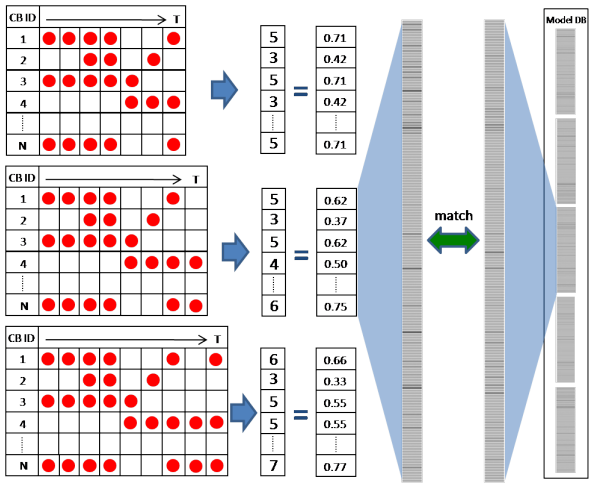

Figure 3. Activation signatures are generated by counting the activations of codebook entries during tracking and normalizing with hypothesis lifetime. The activation signature (visualized by a strip) of a tracked person is matched with the activation signatures in a database of known persons to reject persons that are ineligible.

codebook entries as indexes, we have a structure for feature alignment in person model matching. To make the matching approach even faster, we use a two stage matching strategy which can discard models with a low similarity in the first stage, based on the activation signature only, without matching the feature descriptors. The activation signature thereby is the time normalized activation count of codebook entries and is built as shown in Figure 3. The use of the activation signature allows for faster matching since no high dimensional feature descriptors have to be matched, but only one vector of the codebook dimension N for each person model. In this matching stage, we can spot rough differences between two person but are not able to distinguish persons that activate the same codebook structures. This finer distinction can be accomplished in the second stage where the descriptor clusters of the model entries are compared. Here, differences in the characteristics of certain structures are to be spotted.

For stage 1 matching, only the activation signatures of person models are relevant. These signatures can be directly inferred from the models that were built during person tracking by counting the number of activations per model index as shown in Figure 3. To be independent of the time interval that a person is tracked, the feature count is normalized with the duration of model generation (the time the person was tracked). To match two person models ζ and η, the time normalized activation vectors are compared:

$$\delta_{ACA}(\zeta,\eta) = \frac{1}{N}\sum_{n=0}^{N} \left| \frac{|\zeta_n|}{\zeta^T} - \frac{|\eta_n|}{\eta^T} \right|. \qquad (1)$$

Where η_n is the number of activations of codebook entry n, η^T is the lifetime of hypothesis η (same for ζ), and N is the codebook dimension.

Models can be discarded in this step by application of a threshold to the distance δ_{ACA}. Since the same activation structure is a prerequisite for a high match between two person models, all models that are discarded in this stage could not have gained a high match in the next stage. This first stage can only discard person models which have strong structural differences to the input model. For instance due to different clothing (a person wearing a skirt opposed to a person wearing pants) which leads to local differences in person shape. With an increasing number of persons in the database, we cannot expect the structural differences alone to be sufficient to distinguish people since people might where the same type of clothes. For that reason, the capability to distinguish people the appearance of which only differs little is necessary.

To spot these detail differences between persons, the feature descriptors themselves have to be compared. For that, the complete models including the descriptor clusters are matched, again indexed by the codebook structure which was used in model generation too. By that indexing, we reduce the amount of data that has to be matched because only the descriptors in the same model entry have to be matched. Since these activated the same codebook prototype in object detection, we can expect them to represent the same part of a person (e.g. the head) in most cases. But, since the codebook was built only based on appearance information and spatial similarity (in terms of object center offset) is not demanded in a codebook entry, image features activated by the same codebook entry might in fact refer to different ob-

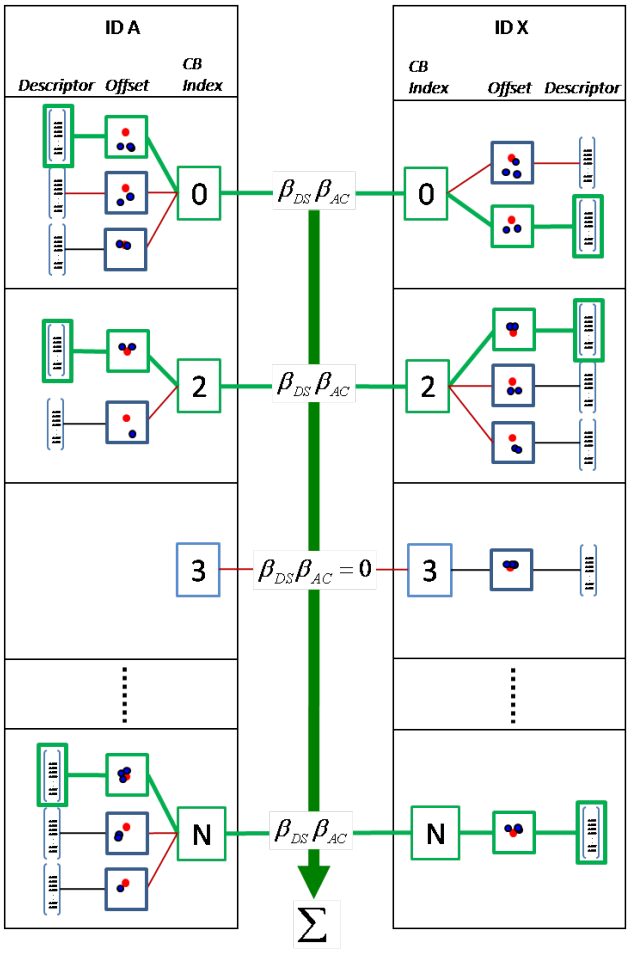

Figure 4. Matching of person models. Models are aligned based on codebook entry indexing. Spatial consistency of features is checked by matching the spatial distributions (object center offsets = blue dots, object center = red dots). Feature descriptors which pass the spatial consistency check are matched. For each model entry, the minimum descriptor distance is picked.

ject parts in some cases. This might happen since different body parts of person might look very alike (like arm and leg), specifically when observed at a very low resolution. To disallow matches between these components, spatial (in terms of object center offset) match of feature descriptors is demanded in addition to descriptor match. This ensures spatial consistency of matching features and additionally reduces the amount of data that has to be processed.

The principal matching strategy is depicted in Figure 4. Models are first aligned by their codebook activation indexes. As we see, feature descriptor matching is conducted only if their distributions have a spatial match. Green lines indicate matches between the two models. The match between two models is the weighted (with the activation match) sum of all model entry descriptor matches where each model entry is counted with its best match. The match sum is further divided by a normalization constant that includes the length of the time interval during that the model was built.

In detail, the overall match β between an unknown person model ζ and a model in the database η is determined by:

$$\beta(\zeta,\eta) = \frac{\sum_{n=0}^{N}(\beta_{DS}(\zeta_n,\eta_n)\cdot\beta_{AC}(\zeta_n,\eta_n))}{\phi(\zeta,\eta)}. \quad (2)$$

With $\beta_{DS}(\zeta_n,\eta_n)$ being the descriptor match for model entry n:

$$\beta_{DS}(\tau,\xi) = \delta_{DS}^{MAX} - min(\delta_{DS}(\tau,\xi),\delta_{DS}^{MAX}). \quad (3)$$

Here, the minimum of the model entry distance $\delta_{DS}(\tau,\xi)$ and a model entry distance maximum threshold δ_{DS}^{MAX} is picked and subtracted from the maximum model distance. By using this measure, we are able to transform distance into similarity and additionally account for the quality of a match. Since all distances above the threshold are zeroed in the match measure, they do not have any influence on the overall match. Here, the model entry distance δ_{DS} is the minimum distance when considering all combinations of descriptors in a model entry (all descriptors in ζ_n are matched with all descriptors in η_n):

$$\delta_{DS}(\zeta_n,\eta_n) = \min_{k,i}(\delta_S(\zeta_{n,i},\eta_{n,k})\cdot\delta_D(\zeta_{n,i},\eta_{n,k})). \quad (4)$$

Whereat δ_D is the descriptor match - we use the Sum of Squared Differences (SSD) for SIFT matching - and δ_S is the spatial match of the feature distributions:

$$\delta_S(\tau,\xi) = \begin{cases} 1, & \text{if } \min_{i,k}(dist_{eukl}(\tau_i,\xi_k)) < \delta_S^{MAX} \\ \infty, & \text{else} \end{cases}. \quad (5)$$

$\beta_{AC}(\zeta_n,\eta_n)$ (in 2) is the activation signature match for model entry n:

$$\beta_{AC}(\zeta_n,\eta_n) = 1.0 - |\frac{|\zeta_n|}{\zeta^T} - \frac{|\eta_n|}{\eta^T}|, \quad (6)$$

and $\phi(\zeta,\eta)$ is the normalization constant that accounts for the model generation duration ζ^T, the normalized sum of activation weights $\sum_{n=0}^{N}\frac{|\eta_n|}{\eta^T}$ (or ζ respectively) and the codebook length N:

$$\phi(\zeta,\eta) = \frac{\zeta^T}{(\frac{1}{N}\sum_{n=0}^{N}\frac{|\zeta_n|}{\zeta^T})(\frac{1}{N}\sum_{n=0}^{N}\frac{|\eta_n|}{\eta^T}))}. \quad (7)$$

This normalization constant has the main effect, that matching is most independent of the duration persons are tracked for model building and the number of features which are acquired during tracking.

Note that this model matching approach is applicable not only for models built during tracking but also for snapshots (single detection) of persons. The problem in that case certainly is, that the articulation of a person has a great deal of influence on the matching result.

4. Evaluation

We evaluate our person reidentification approach on a subset of the infrared dataset (dataset C) of the CASIA Gait Database [1]. This dataset was originally generated for gait recognition purposes but it is perfectly suited for person reidentification evaluation because is consists of annotated short sequences of different persons. Multiple sequences exists for each person which suits it for reidentification purposes because we need at least two sequences of the same person for evaluation. For this evaluation, we pick 15 different persons to build a database as basis for reidentification. For person detection, a detector trained for infrared persons is applied. Person models are built during tracking with the approach described in section 2 without any artificially generated training sets. Due to possible imperfections in tracking, the tracking, and thus the model generation duration might vary for different persons. This offset is typical for real application scenarios where people stay in the observed scene for different durations. As we depicted in section 3, our matching model deals with these real world problems by using normalization factors. For reidentification evaluation, we pick a second sequence for each of the 15 persons in the database and a single sequence for 5 additional persons which are not in the database. As we see in Figure 5 (a) and (b), these people all look very alike and are, even for a human being, difficult to distinguish. Using these 20 sequences, we perform an open-set classification. Open-set in this context means, that not all person that enter the scene are in the database. For reidentification, this means that it is to decide if a person has been seen before (is in the database) and, if true, which person in the database is our current person. This simulates a usual reidentification task in surveillance scenarios where an unknown number of people enter, leave and reenter the scene.

From that task definition, three corresponding error rates can be derived. The *false rejection rate (FRR)* is the rate of persons which could not be reidentified but actually are in the database (so falsely classified as unknown):

$$FFR = \frac{\#false\ rejections}{\#known\ samples}. \tag{8}$$

The *false acceptance rate (FAR)* is the rate of persons which are accepted as known persons (the system has reidentified the currently tracked person as a person inside the database), but actually have not been seen before:

$$FAR = \frac{\#false\ acceptances}{\#unknown\ samples}, \tag{9}$$

(a) *Persons in the database.*

(b) *Persons that are not in the database and thus are to be classified as unknown.*

Figure 5. Example images of the dataset used for testing.

and the *misclassification rate (MCR)* is the rate of persons which are identified as the wrong person:

$$MCR = \frac{\#misclassifications}{\#known\ samples}. \tag{10}$$

A correctness measure that joins FAR and MCR is the *correct classification rate*:

$$CCR = 1.0 - MCR - FRR \tag{11}$$

$$= \frac{\#correct\ classifications}{\#known\ samples}. \tag{12}$$

The distance measure used for matching a tracked person with models in the database was introduced in section 3.2. Although this distance measure is most independent of the time a person is tracked and the number of features a model includes, it is difficult to classify people based on an absolute match value since different people have different characteristics which lead to differences in the number of features which are found on a person in principle. This principal offset for certain persons is (due to our distance measure) uncritical regarding the relative distances to other

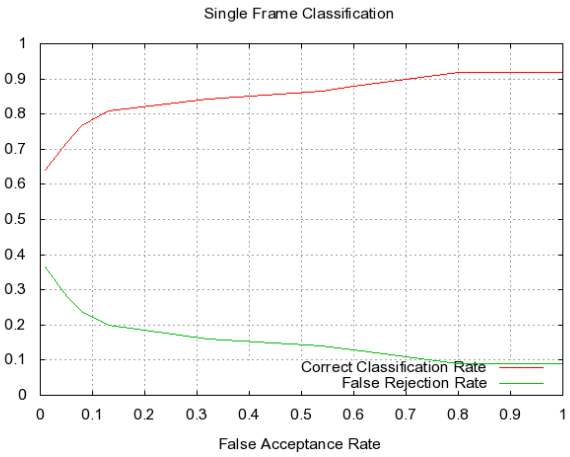

Figure 6. Single frame classification rates.

persons, but it makes the choice of a global threshold difficult, specifically in an open-set classification task where this threshold is used for rejecting persons and picking the correct person. For that reason, we base classification on the ratio ρ of best match and second best match of the model that has to be identified to models in the database. By that, classification is independent of the absolute match and thus of possible offsets induced by person characteristics or signal level issues.

The decision, whether a tracked person is classified as a certain database entry or rejected as "not seen before" (in surveillance tasks, a new model would be stored in the database for this person) is based on ρ. If ρ is above a threshold we can classify the tracked person as the best matching database entry. Since our approach is able to provide classification results at any time during a person is tracked, we do not have to know the whole sequence before we are able to provide a result. Our approach can thus provide a classification result for the first frame, based only on a single perception of the person and for the last frame of the sequence based on the information of the whole sequence. To regard these two aspect of our system, we first evaluate per frame classification and then provide classification results per sequence.

Per frame classification results are shown in Figure 6. The plot was generated by applying different thresholds for classification. It shows CCR and FRR as a function of FAR. As we see, good results are accomplished even when considering each track result separately. Another details is, that FRR is the mirrored (at the shifted x axis) CCR. This is because MCR is infinitesimal small (and thus not plotted here) because we have only 4 misclassifications in total for all 15 sequences, which each has about 100 frames. This means, the correct person in the database nearly always (except 4 times for a single person) has the best match. The goal thus mainly is to find a good threshold that separates people in the database from those not in the database.

Single frame classification is an uncommon evaluation method in this context since tracking provides us with a time series of results and the decision only has to be made for the whole track of a person once. We thus evaluate classification when considering the whole time series. For that, we analyze different combinations of ρ thresholds and temporal consistency demands. The maximum classification correctness is reached at a threshold of 2.2 for ρ and a temporal consistency demand of 20%. This means, that a tracked person is classified as a certain person in the database if ρ exceeds a threshold of 2.2 20% of the tracked time for this certain database entry. (In addition, this database entry should be the best database match a minimum of 51% of the track duration. This is always met in our experiments, since we have only 4 frames misclassification on the total set). Using this classification criterion, our reidentification system has a correct classification rate of 95% which means 14 out of 15 known people are correctly reidentified and 5 of 5 unknown people are correctly classified as unknown. This is shown in detail in Figure 7, which shows the per frame correct classification rates for every ID separately. We see, that the temporal cosistency demand of 20% is exceeded by far by most persons.

The results for single frame classification (threshold picked for Equal Error Rate (EER)) and sequence classification are summarized in Table 1.

5. Conclusion

In this paper, we presented a person reidentification approach for infrared image sequences. For that, we introduced a local feature based combined person detection, tracking and reidentification strategy that is, as a whole, completely self contained and thus applicable independently of specific application scenarios. For person reidentification, we introduced a novel model that uses the general appearance codebook applied for person detection as indexing structure for reidentification and thus is able to efficiently acquire and match models. For matching models we developed a two staged strategy, that allows for fast discovery of promising models based on matching of person signatures in the first stage and detailed analysis of models in the second stage based on feature descriptors. We evaluated the reidentification approach in a subset of the CASIA infrared dataset. The results show good performance when

	FAR	FRR	MCR	**CCR**
Single frame classification	0.18	0.18	0	**0.81**
Sequence classification	0	0.05	0	**0.95**

Table 1. Reidentification results.

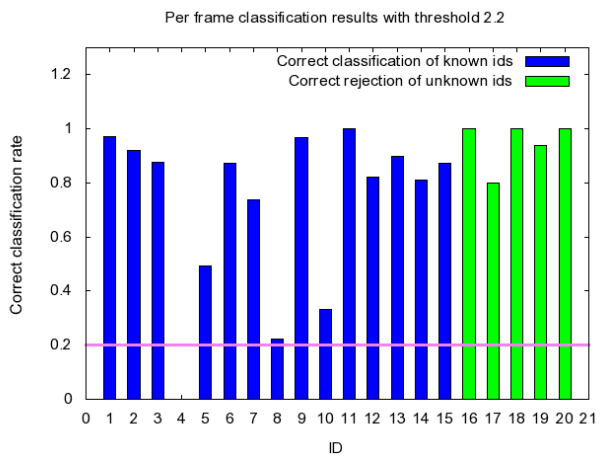

Figure 7. Correct classification rates for each person separately using a threshold of 2.2 for ρ. X-axis shows the different persons, y-axis shows the time slice (e.g. 0.5 means 50% of the hypothesis lifetime) the person is correctly classified (accepted as the correct person or rejected as unknown) when demanding a minimum ρ of 2.2. We can see, that with a temporal consistency demand of 20% (0.2), 19 out of 20 persons are classified correctly.

classifying on single frame basis and nearly perfect performance in image sequence classification. This good performance is remarkable since people look very alike in infrared and are difficult to reidentify even for a lifelong trained human being. However, the sequences used for evaluation have the major advantage that people are visible only in side view and multiple articulation are observed in an image sequence. In real world scenarios this might not always be the case. Consequently, our future work will include the analysis of articulation influence in person reidentification and thereby mainly how to identify local features that are best suited for person reidentification. Here, our plans include the use of part semantics to identify promising features.

6. Acknowledgements

Portions of the research in this paper use the CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences

References

[1] Casia gait database, http://www.sinobiometrics.com. obtained from http://www.cbsr.ia.ac.cn/english/gait6

[2] C. Arth, C. Leistner, and H. Bischof. Object reacquisition and tracking in large-scale smart camera networks. In *ACM/IEEE Int. Conference on Distributed Smart Cameras, 2007. ICDSC '07*, pages 156–163. IEEE, September 2007. 2

[3] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Proc. 9th European Conference on Computer Vision*, pages 404–417, Graz, Austria, May 2006. 1

[4] C. Dai, Y. Zheng, and X. Li. Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computational Visual Image Understanding*, 106(2-3):288–299, 2007. 1

[5] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. volume 2, pages 1528–1535, Los Alamitos, CA, USA, 2006. IEEE Computer Society. 1

[6] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ACM/IEEE Int. Conference on Distributed Smart Cameras*, pages 1–6, September 2008. 1

[7] G. Jaffre and P. Joly. Costume: A new feature for automatic video content indexing. In *Proc. of RIAO*, page 314325, 2004. 1

[8] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society. 1

[9] G. T. Junfeng Ge, Yupin Luo. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *Intelligent Transportation Systems, IEEE Transactions on*, 10(2):283 –298, June 2009. 1

[10] K. Jüngling and M. Arens. Detection and tracking of objects with direct integration of perception and expectation. In *Proc. Int. Conference on Computer Vision, ICCV Workshops*, pages 1129–1136, 2009. 2

[11] K. Jüngling and M. Arens. Feature based person detection beyond the visible spectrum. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, pages 30–37, 2009. 2, 3

[12] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. Journal of Computer Vision*, 77:259–289, 2008. 2

[13] A. Leykin and R. Hammoud. Robust multi-pedestrian tracking in thermal-visible surveillance videos. In *Proc. Conference on Computer Vision and Pattern Recognition Workshop*, page 136, June 17–22, 2006. 1

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

[15] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise:visual search engine using multiple networked cameras. In *Proc. of the IEEE Conference on Pattern Recognition*, pages 1204–1207, August 2006. 1

[16] F. Porikli. Inter-camera color calibration by correlation model function. In *Proc. of IEEE International Conference on Image Processing*, pages 133–136, 2003. 1

[17] F. Tang and H. Tao. Object tracking with dynamic feature graph. In *IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 25–32. IEEE, October 2005. 2

[18] F. Xu and K. Fujimura. Pedestrian detection and tracking with night vision. In *Proc. IEEE Intelligent Vehicle Symposium*, volume 1, pages 21–30, June 17–21, 2002. 1