in: Proc. 9th IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS 2012), Beijing, China, pp. 184-190, IEEE 2012

View Invariant Appearance-based Person Reidentification Using Fast Online Feature Selection and Score Level Fusion

Markus Eisenbach, Alexander Kolarow, Konrad Schenk, Klaus Debes, and Horst-Michael Gross*
Neuroinformatics and Cognitive Robotics Lab
Ilmenau University of Technology
98684 Ilmenau, Germany
markus.eisenbach@tu-ilmenau.de

Abstract

Fast and robust person reidentification is an important task in multi-camera surveillance and automated access control. We present an efficient appearance-based algorithm, able to reidentify a person regardless of occlusions, distance to the camera, and changes in view and lighting. The use of fast online feature selection techniques enables us to perform reidentification in hyper-real-time for a multi-camera system, by taking only 10 seconds for evaluating 100 minutes of HD-video data. We demonstrate, that our approach surpasses current appearance-based state-of-the-art in reidentification quality and computational speed and sets a new reference in non-biometric reidentification.

1. Introduction

Fast and robust person reidentification is a key condition for multi-camera surveillance applications. Tasks vary from supporting image-based tracking in ambiguous situations, cross-camera global tracking, up to automated access control for restricted areas.

In this paper, we focus on cross-camera global person tracking in an uncontrolled multi-camera surveillance scenario, like the terminal building of an airport. Common reidentification approaches from biometrics (face, iris, fingerprint, and gait) are known to be very robust for controlled scenarios (e.g. automatic passport control), with cooperating passengers. In uncontrolled scenarios, biometric approaches are not (iris, fingerprint) or only limited applicable (face, gait). Facial person reidentification is easily avoided by turning away from the camera or occluding the face. Gait recognition often fails due to missing full body views. Therefore, appearance based reidentification methods are preferable for uncontrolled scenarios. Nevertheless, face and gait-based reidentification can be used in parallel to this.

To realize a reidentification system for uncontrolled surveillance scenarios, it is often not necessary to distinguish all persons from each other, but to only reidentify one selected person. In those systems, it is sufficient to use the selected person's model and rank all persons according to their similarity. In comparison to binary classification, a ranking is often better suited for reidentification tasks, since it provides additional score (confidence) values. Therefore, we introduce a novel reidentification approach that provides scores and rankings, while being real-time capable for a multi-camera surveillance system.

We focus on the scenario of airport surveillance. Our reidentification method supports an operator in finding a selected person. All search tasks are triggered by the operator. For speeding up the operator triggered search, features for reidentification are precalculated for all persons during the recording of the videos.

The remainder of the paper is organized as follows: In Sect. 2, we summarize current state-of-the-art methods for person reidentification. Then, we introduce the proposed reidentification method in Sect. 3. In Sect. 4, we evaluate our approach on a public benchmark dataset and a real-world surveillance application. We end with a conclusion.

2. State-of-the-Art

Recently, a lot of research is done in the field of person reidentification. A coarse, but systematic review of the different approaches is given in Fig. 1. The state-of-the-art differs in the capabilities concerning a variety of challenges. A comparison of some of the most important algorithms with the method proposed here is given in Tab. 1. For

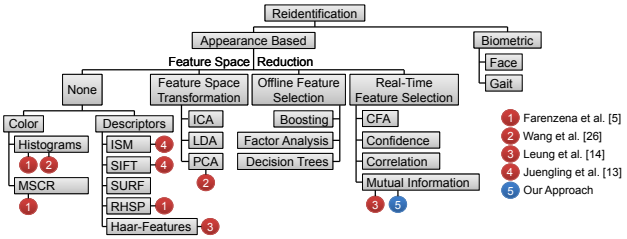


Figure 1. Categorization of reidentification methods

*This work has received funding from the German Federal Ministry of Education and Research as part of the APFel project under grant agreement no. 13N10797.

in: Proc. 9th IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS 2012), Beijing, China, pp. 184-190, IEEE 2012

classification, our approach is applicable for view-invariant appearance-based reidentification with occlusions present. It uses a feature selection for feature space reduction. Due to the reduced feature set, it is real-time capable in the execution phase.

Method	VIA	FSR	FS	OCC	RTC	SPS	NNS	FUS
MSCR, RHSP [5]	×	×	×	×		×	×	
Histogr., PCA [26]	×	×		×	×			
Haar-Ft., MI [14]	×	×	×		×			
SIFT, ISM [13]	×			×		×	×	
Proposed Method	×	×	×	×	×			×

Table 1. Comparison of state-of-the-art reidentification approaches. Columns indicate, whether the denoted methods can handle view-invariant appearances (VIA), do a feature space reduction (FSR), use feature selection (FS), can handle occlusions (OCC), are real-time capable (RTC), can be trained on a single positive sample (SPS), do not need negative samples for training (NNS), and are easily applicable for fusion (FUS).

3. Ranking-based Reidentification Using Fast Online Feature Selection

In this section, we present our reidentification method which provides similarity scores and rankings instead of only binary decisions. Ranking person hypotheses has several advantages compared to strict classification. While classification only produces a definite response, the ranking can be used even in ambiguous situations by providing scores for each hypothesis. As for example, in a situation, where no hypothesis can be accepted due to ambiguities, our method can at least reject all hypotheses, that are definitely unsuitable, without the risk of rejecting the person of interest.

A tremendous amount of different features can be extracted from images in order to realize appearance-based person reidentification. Depending to the current situation in the scene, different sets of features are suited to realize a robust reidentification. Our approach uses feature selection and information-theoretic learning, to find a set of suitable features to circumscribe one person compared to all other persons in the scene. Afterwards, the ranking provides the similarity of a person's model to all other persons.

To establish a ranking, we have to choose a person representation that can be used to provide a similarity function. For reasons of simplicity, we use a multi dimensional mixture of Gaussian probability distribution [2]. To estimate the similarity of a person hypothesis to a learned person specific model, we compute the Mahalanobis distance. Using the ascertained distances for all persons in the test dataset, a ranking can be established. A shortcoming of the chosen approach is the need for estimating high dimensional covariance matrices, which cannot be approximated with a small number of samples. In surveillance scenarios, usually only 100-500 samples (several seconds of video sequences) are available for model training. Therefore, this model cannot be applied to high dimensional feature spaces in such a scenario. To overcome this "curse of dimensionality" dilemma in person reidentification, most approaches use a preselected set of features to reduce the dimension of

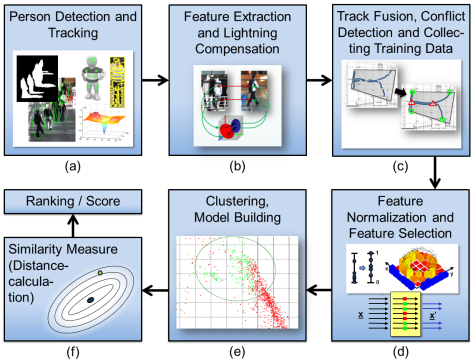


Figure 2. Proposed algorithm, composed of live analysis (a-b), model training (c-e), and execution phase (f).

the feature space. This is counterproductive, since without the knowledge of the current scene and the persons within the scene, only a sufficiently manifold feature space makes it possible to separate the persons from each other. Choosing a small preselected set of features could fail in cases with many persons, since the relevant features for separating this person from others might be left out.

The approach we introduce uses online feature selection to find a discriminative but small feature set, which can be used for the selected person, rather than separating all persons from each other. This enables us to estimate the covariance matrix for a person specific similarity function at runtime. Containing only the most discriminative features, this can be achieved even with relatively few training samples.

Fig. 2 shows all steps of our approach: The first step is the detection and tracking of all persons in the scene. The features, needed for reidentification, are extracted for each tracking hypothesis. This "live analysis phase" is executed in parallel to the recording of the video sequences (Sect. 3.1).

After a person is selected, the "model training phase" starts. Model training cannot be applied for all persons in the scene, since it is computationally very expensive. For model training, a dataset is generated using well suited tracks of the selected person and others in the scene. Then, the features of the training data are normalized, and online feature selection is performed using the joint mutual information [28] for subsets of features. After the features were selected, the space spanned by the remaining features is clustered to get the person specific multi-Gaussian model (Sect. 3.2).

Having all models trained, the score calculation and ranking can be done very fast. Score-level-fusion is used to increase the performance. The start of this "execution phase" is triggered by an operator (Sect. 3.3).

3.1. Live Analysis Phase

Real-Time Person Detection and Tracking

In order to reduce the number of reidentification tasks, robust components for person detection and tracking are needed to produce long and non-ambiguous trajectories. In our approach, real-time person detection is done using an advanced version of contour cues [27], which is speeded

up by foreground segmentation and calibrated cameras. For person tracking, we implemented a simple and fast sparse template-based feature tracker. We improved the approach of [6], by using color features drawn from homogeneous regions and choosing a faster direct search strategy (logarithmic search [10]). This template-based approach has the advantage that hypotheses with occlusions can be detected and handled separately in the model training phase. We decided in favor of these person detection and tracking methods, since they perform at high frame rates, leaving enough computational time to extract lots of features. Nevertheless, these components are exchangeable, and thus, the algorithm can easily be adapted to other applications.

Feature Extraction

Basically, all kind of features can be used for this appearance-based reidentification approach. But since feature selection should only select a minimum number of most relevant feature space dimensions, one single feature should only consist of one channel, or a small number of channels, which can be decomposed into a set of discriminative channels. We characterize a channel as a single component of a set of elements assembling a feature. As example, we compare the channels of RGB and SIFT:

- The RGB-mean-color feature consists of three decomposable discriminative channels (red, green and blue), which makes this feature applicable.
- A SIFT-descriptor has a set of 128 channels, each alone without discriminative power. Therefore, this descriptor can only be applied, if it is transformed into a single channel using a similarity function.

A second condition for the features is, that they have to be extracted in real-time for the proposed surveillance scenario. Fig. 3 shows a categorization of possible features.

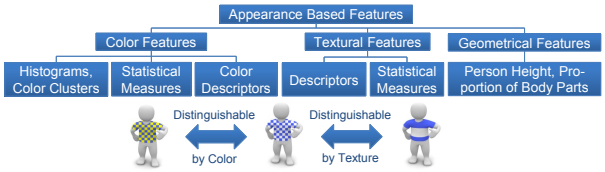


Figure 3. Categorization of features suitable for appearance-based reidentification

At first, to find the most relevant features for a particular reidentification task, we extract as much features as possible during the live analysis phase. To emphasize the recognition capabilities of our approach, we restrict the feature set in in this paper's experiments as follows:

- 13 texture features from Haralick et al. [8] ($f_1 - f_{13}$).
- The mean color of a defined region in 9 different color spaces [4], [24] (RGB, YC_bC_r, HSV, HSL, HSI, RG-BY-WS [19], XYZ, CIE $L*a*b*$, $I_1I_2I_3$ [17])

The features are extracted from two predefined regions (upper and lower body) using the tracking hypotheses (see Fig. 4). These two positions were experimentally determined to be very robust for matching foreground, while being stable in color and texture for different views and perspectives (including occlusions by body parts and diverse clothing). Again, this is done for efficiency. An exact segmentation of the person would eventually produce more stable regions, which may increase the performance. If our approach is adapted to other tasks, like animal or car reidentification, the segmentation step could even be obligatory if the position of the regions would change for different viewpoints.

Figure 4. Regions for feature extraction. The left drawing shows the relative position of the used regions on upper and lower body to the tracking hypothesis. On the right, sample images for different persons and views, and the extracted regions are shown.

3.2. Model Training Phase

Training Dataset Generation

The quality of our real-time reidentification system greatly depends on the quality of the training dataset. For a good training dataset, it is essential to have enough non-ambiguous training samples with a high variety of different perspectives.

Since the features for reidentification are extracted automatically, the system needs to resolve possible conflicts like occlusions and ID-switches by its own, while providing a high variety of samples, like perspective and lighting changes. In scenes with multiple overlapping cameras, sensor fusion techniques can be used to build long continuous tracks using geometrical dependencies [25]. Using a network of calibrated cameras, the persons are tracked in global coordinates. In regions with overlap, sensor fusion techniques [16] can be used to resolve spatial temporal dependencies, in order to connect tracks from different cameras. Using tracks from more than one camera provides a greater variety of different perspectives for training.

Global tracking can also be helpful to avoid possible conflicts while building the training dataset. Occlusions between persons and the risk of ID-switches can be detected by taking the geometrical constraints between the persons and the camera into account. To avoid false samples, these conflict situations are excluded from the training.

Feature Selection based on Information-theoretic Learning

In this subsection, we will introduce the online feature selection, which is used to reduce the dimensions of the feature space. At the end of this subsection, we also discuss the dimension reduction via subspaces, like Principal Component Analysis (PCA).

For feature selection, the mutual information (MI; see [20]) can be used to approximate the dependence between a feature channel and the corresponding class labels (negative/positive). To evaluate the mutual dependence between a set of channels and the class labels, the joint mutual information is used (JMI; see [28]).

In [20], methods for estimating the MI for feature selection are compared. Their experiments conclude that using the correct ranking of the MI is sufficient for feature selection. This makes a correct calculation of the MI obsolete. The experiments also show that a correct ranking is already achieved with simple histogram based methods. Since simplicity beats complexity, we use a simple histogram approach with equal bin width. This reduces the complexity of the implementation of MI and JMI, since the typical integrals are replaced with sums and the probability distributions are represented by histograms. For further details, it is referred to [23]. Our approach additionally includes a weighting of samples and an exceptional handling of multimodal distributions. This increased the performance significantly.

The following two approaches are common for selecting the correct features:

- MI in combination with the MIFS-algorithm [1]
- computing the JMI for different subsets of features and choosing the subsets with the highest JMI

Our evaluation shows that MIFS performs badly on our datasets. We believe that this is caused by the choice of the first channel, which has a large influence on the choices of the remaining features. Since MIFS only uses the MI, it is deficient if XOR problems are present. In XOR problems, the MI of both channels is 0, even though the JMI is 1. Since we detected XOR like problems in feature spaces for reidentification, we decided in favor of the JMI. Nevertheless, also the JMI-based feature selection has some shortcomings:

- **High dimensional distributions:** Trying to approximate a high dimensional distribution with only few samples is problematic. Therefore, only small sets of channels can be considered (in surveillance scenarios 3-6, dependent on the number of samples).
- **Computation time:** Choosing a suitable set of channels is computationally expensive, since every possible combination of channels needs to be evaluated.
- **No incremental selection possible:** Incremental selection of features will fail, since it lacks the same problem as the MI for XOR like problems.

Since it is almost impossible to compute the JMI for all combinations of channels, we divide the channels into nearly independent groups (five to ten channels). This can be achieved by using the MI or designer knowledge. Using small groups, the best subset of channels, based on the JMI, is chosen for each group. In the next iterations, close groups are combined to find new subsets and non selected channels are removed. The elimination of channels is repeated until a maximum number of channels is selected.

Also a transformation to a subspaces would be applicable to reduce the number of dimensions. But most techniques of this category are unsupervised (Principle Component Analysis (PCA), Independent Component Analysis (ICA), Non-Negative Matrix Factorization (NMF) [7]). This means, they ignore the class labels, which causes a high risk of eliminating important dimensions for separating two persons. Therefore, the use of these unsupervised methods may lead to errors and should be avoided. Otherwise, the most common supervised method, Linear Discriminant Analysis (LDA) [7], is only applicable for linear separable data, and therefore, it is not preferable for our scenario.

Building the Similarity Function

Having the N most significant features selected, the next step is straight forward: In the N-dimensional feature space, we utilize the mean shift algorithm with a Gaussian kernel to cluster the positive samples. For each cluster, we calculate the mean vector and covariance matrix. This creates the multi-dimensional Mixture-of-Gaussian model and enables us to use the Mahalanobis distance to the nearest Gaussian kernel as similarity measure.

The model training phase takes 2 seconds (40 channels, 10 seconds of video data for training) up to 40 minutes (600 channels, 1 hour of video data for training), strongly depending on the size of the training data set and the number of features. In surveillance scenarios, usually the time needed for model training does not exceed 2 minutes (for 600 channels) due to small training data sets. This can easily be compensated in the execution phase (see below). When dealing with huge training data sets, subsampling should be performed to reduce the time needed for training.

3.3. Execution Phase

Scores and Rankings

In the execution phase, which is dealing with the person search triggered by an operator, the model of the selected person is used for reidentification. To calculate the similarity d_i of a test sample i to the model's nearest of k Gaussian kernels, we use the Mahalanobis distance d_{ij} as similarity measure ($d_i = \operatorname{argmin}_j (d_{ij}), j = 1 \ldots k$). Since only using a single sample is not very robust, we calculate a score s for all spatio-temporally associated observations of a person (termed "track"), based on the score of each sample of the track.

For track comparison, the benefits of using a similarity measure instead of a classification become apparent: Instead of using a majority vote, we can use the best match $s_{best} = \operatorname{argmin}_i (d_i)$, the average distance $s_{avg} = \operatorname{avg}_i (d_i)$, or a combination of both. Our experiments show, that using a combination is the best choice ($s_{combi} = \operatorname{argmin}_i (d_i) + \operatorname{avg}_i (d_i)$). Having the scores for all candidates calculated, we construct a ranking, based on the score values. For deciding which person fits best to the model, we use the following criteria:

- For closed set scenarios (e.g. supporting image-based tracking in ambiguous situations), we choose the person ranked first, if the difference between the scores of the first and second rank is large enough. Otherwise, we use score level fusion with other reidentification methods (see below).
- For open set scenarios (e.g. multi-camera surveillance systems), we choose the person ranked first, if the score is better than a global threshold. Otherwise, we use score level fusion with another method, see below again.

Due to the reduced feature set and the use of a very simple similarity measure, the execution phase is very efficient.

We are able to perform 12 000 score calculations within one second on an Intel core i7 system. Therefore, the time for model training can easily be compensated in the execution phase.

Score-Level-Fusion

For decision fusion, a large variety of methods exists [15]. We use a score-level-fusion mechanism, that permits the fusion of our method with any other reidentification approach on an abstract level. To be able to do so, both methods must provide a normalized score and have to be almost statistically independent. This can be achieved by working on different data (e.g. using features from upper and lower body), or using different features (e.g. skin color and a face reidentification mechanism).

A good choice to make the scores of all methods comparable, is the normalization regarding the false acceptance rate (FAR), since the mapping can be learned data-driven. Therefore, we use score values of person comparisons with models of any other person (which would be false positives, if accepted), on a huge dataset with known ground truth (e.g. public available benchmark dataset). Then we map the k collected scores for method m to the FAR (Eq. 1)

$$FAR(s_i^m) = \frac{rank(s_i^m)}{\text{argmax}_j \, rank(s_j^m)} \tag{1}$$

using the ratio between the rank of a score s_i and the maximum rank of all scores $s_j, j = 1 \ldots k$ (= the probability for observing a false positive as a function of the score). Since frequent calculations of ranks in the execution phase would be computationally expensive, a lookup table is used instead. The use of a logarithmic measure (Eq. 2) [9] is preferable to Eq. 1, due to the simplification of the fusion of two scores of methods m and n to a single addition (Eq. 3) which correlates with a multiplication of the two methods' probabilities of accepting a false person (Eq. 4).

$$s_i^{norm,m} = -\log_{10} FAR(s_i^m) \tag{2}$$

$$s^{fus} = s^{norm,m} + s^{norm,n} \tag{3}$$

$$= -\log_{10}(FAR(s^m) \cdot FAR(s^n)) \tag{4}$$

Using normalized scores has an additional major advantage: It can be interpreted easily. E.g. a score $s^{norm} = 2$ depicts a probability for a false acceptance of $1:100$, while a score $s^{norm} = 6$ stands for a probability for a false acceptance of $1:1\,000\,000$. This can be extremely helpful, if a threshold for decision making has to be chosen. As for example, a surveillance system has to identify 1000 persons per hour, and one false decision per hour is considered to be acceptable, a threshold of 3 should be the choice (chance of $1:1000$ for a mismatch).

4. Evaluation

In order to evaluate our approach, we used the Casia A dataset [3] (Fig. 5(a)). It was originally designed for gait recognition, but it has also been used for benchmarking view invariant person reidentification [13]. It contains recordings of 16 persons walking in six different directions in two sequences. We compare our approach to [13], which uses SIFT and ISM features for reidentification. So far, this method performed best on the chosen dataset. Moreover, we demonstrate the performance of our approach on a surveillance system, installed in the terminal building of the Erfurt-Weimar airport, Thuringia, Germany (Fig. 5(b)).

Figure 5. (a) Casia A dataset [3], (b) Airport scenario

View Invariant Person Reidentification

For the Casia A dataset (see Fig. 5(a)), the task is to assign a test sequence of a person to the best fitting model in a database. Therefore, we have to deviate from our normal scheme: Instead of ranking a set of persons for a model, we rank a set of models for a given person.

For the database, we had to train models on all 192 sequences (16 persons, 6 views, 2 sequences per view), resulting in 32 models per view. Since we need both positive and negative samples, we divided the 32 sequences of each view into two parts. Negative samples were only drawn from the 15 negative sequences from the same part as the actual positive sequence. This way we trained all the models.

As [13], we also conducted a closed set evaluation. In the execution phase, all 32 models of a single view are used. The 32 sequences of a second view are used as test samples. For our approach the task is to calculate a ranking for each sequence, based on the similarity to each model. For evaluation, we use the Correct Classification Rate (CCR) [12]. To get a fair comparison, we only assess the ranking as correct, if the two models (out of 32) belonging to the same person as the test sequence are ranked first and second. Every model of the person ranked worse is called a mismatch. For the same view, we leave out the model trained on the test sequence, and evaluate if the remaining model for this person is ranked first (out of 31 models). The evaluation is done for every possible view combination (for training and test phase).

Tab. 2 shows the results of our approach, for different viewpoint combinations of the CASIA A dataset, using only features extracted from the upper body. Tab. 3 depicts the results, obtained by using features extracted from lower body. For comparison, Tab. 4 presents the results of [13].

As it can be seen, the usage of simple but discriminative features leads to very good results for view-invariant person reidentification. The reason can be explained using following example: The color and texture of clothes (used in our approach) are usually the same for different orientations, but a highly discriminative feature as that one used

Angle	0°	90°	135°	180°	270°	315°
0°	100	78	94	100	81	92
90°	73	100	83	77	89	75
135°	84	89	94	83	78	86
180°	100	78	88	100	78	94
270°	81	91	88	81	100	91
315°	86	66	77	78	89	100

Table 2. CCR of our approach with upper body features for different view combinations of the CASIA A dataset. Rows show the view used for model training, and columns the view of test samples. For easy comparison to the best approach so far presented in [13], the results are highlighted green for better, black for equal, and red for worse performance.

Angle	0°	90°	135°	180°	270°	315°
0°	94	70	72	86	72	75
90°	50	100	64	45	94	61
135°	80	89	88	70	83	88
180°	89	72	73	91	70	75
270°	61	94	75	52	100	80
315°	83	66	80	66	73	94

Table 3. CCR of our appr. with lower body feat., highlighted as in Tab. 2.

Angle	0°	90°	135°	180°	270°	315°
0°	93	25	42	81	27	57
90°	25	100	36	20	67	34
135°	42	36	100	50	36	72
180°	81	20	50	93	20	25
270°	27	67	36	20	100	42
315°	57	34	72	25	42	100

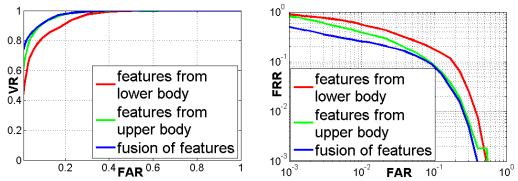
Table 4. CCR of [13] for CASIA A dataset.

in [13] can only be found on a preferred direction. Therefore, a color or texture feature sampled from one direction is more likely to be seen in another orientation. Although the results obtained so far are very good, they can be improved further by not using only features from one region. Therefore, in our third experiment we evaluate the score-level-fusion component of our approach for fusion of the reidentification methods using only upper body respectively lower body features. As shown in Tab. 5, the performance can be increased significantly. We outperform [13] in every viewpoint combination considerably. If the views for training and execution phase resemble, we reach a perfect classification. All scores for same views in training and execution phase were suitable for being accepted in an open set scenario. The lowest score was $s^{norm} = 2.1$, which gives a probability for a mismatch of only $\frac{1}{126}$. Fig. 6 shows the ROC and DET curve for scores of all viewpoint combinations on the Casia A dataset.

Since the performance increase for reidentification using score-level-fusion appears to be minor in Fig. 6 in comparison to the reidentification using upper body features, the FAR-axis is shown logarithmically scaled in Fig. 7. The most significant region for a good rank is within the range of a FAR of 10^{-2} and lower. The much higher verification rate of the fusion in this region in comparison to the single methods is obvious.

Angle	0°	90°	135°	180°	270°	315°
0°	100	91	100	100	92	98
90°	78	100	89	77	95	78
135°	98	98	100	97	97	94
180°	100	94	100	100	91	97
270°	88	100	98	86	100	100
315°	97	94	100	94	95	100

Table 5. CCR for score-level-fusion of reidentification results with upper and lower body features, highlighted as in Tab. 2.



(a) ROC curve (b) DET curve

Figure 6. Receiver Operating Characteristic (Verification Rate, False Acceptance Rate), (b) Detection Error Tradeoff (False Rejection Rate, False Acceptance Rate) on Casia A dataset.

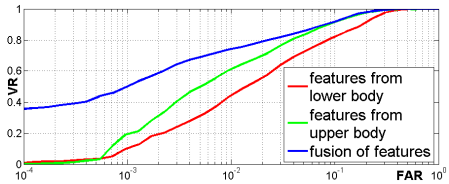


Figure 7. ROC curve of Fig. 6(a) with logarithmic scaled FAR-axis.

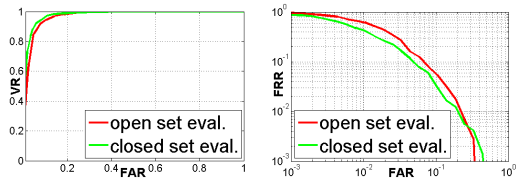
Application on a Real Surveillance Scenario

Evaluating our reidentification approach only on an artificial dataset does not take the importance of training dataset generation into account. This training dataset generation component chooses suitable negative samples and rejects unsuitable training data, and therefore is essential for the quality improvement of the training data set. Therefore, we evaluated our approach, including all components, on a realistic surveillance scenario. With two non-overlapping cameras, we recorded 50 minutes of HD-video data with 10fps in the terminal building of the Erfurt-Weimar airport, Germany.

The scenario includes frequent occlusions, changes in lighting, and varying views. Eight persons traversed the scenes frequently. The task was to mark every occurrence of a person selected by an operator on the whole recording of both cameras.

The Ground-Truth, that was needed to assign person hypotheses between the non overlapping cameras, was extracted with an automatically calibrated Laser-Range-Finder-network [22], that covered the view area of both cameras and the space in between. We used five LRF at the height of 0.7 m and the tracking algorithm described in [21] to record the movement paths of all persons in the scene. The foot points were projected into the calibrated cameras (spatio-temporally synchronized with the LRF-Network) and the ROIs were approximated using an constant person height and height/width-ratio.

The visual person tracker extracted 1 562 tracks (average track length 48 frames, person height 150px–1000px). The training-dataset-generation component collected various different views of the selected persons and negative samples, which were well suited for model training. Fig. 8 shows the ROC and DET curves for a closed set (only the eight persons in the test set) and an open set evaluation (all persons on the airport) for all scenes of both cameras. More than 80% of false hypotheses could be rejected early, due to a very low similarity to the model of the selected person. In unambiguous situations, the surveillance system reidenti-



(a) ROC curve (b) DET curve

Figure 8. Receiver Operating Characteristic (Verification Rate, False Acceptance Rate) and Detection Error Tradeoff (False Rejection Rate, False Acceptance Rate) on a realistic surveillance scenario.

fied all persons correctly. In some cases, persons not present in the training data, produced high scores, but could still be distinguished from the selected person. Only situations with many occlusions by other persons yield to errors.

Reidentification of one person for all frames on 100 minutes of HD-video data, using precomputed features from live analysis phase during recording, took less then 10 seconds on 4 cores of an Intel core i7.

Although, the proposed method performs very well in our experiments, some failure modes should be mentioned:

The chosen person detector and tracker are only able to handle upright standing persons. If other objects or persons in other poses should be reidentified, the detector and tracker must be replaced.

In the experiments, we use the mean color of a predefined region as feature. This can cause problems when lighting changes appear frequently (e.g. in outdoor scenarios) or the lighting of the training data differs a lot from the test data. There are two possibilities to compensate this problem: (A) Color invariants can be added to the feature set. They will be automatically selected if they are better suited, to separate the person from others, than pure colors. (B) Colors can be normalized with a lighting compensation method. An adapted version of [18] shows promising results in our first experiments. If no lighting compensation is done, the quality of the ranking degrades only slightly. The person of interest will still be ranked in the top 10%.

Another problem is the presence of many (> 100) persons. They cannot be distinguished by their appearance with only some features. But again, the positive samples should be ranked in the top 10% using the proposed method. A cascaded approach, which considers only the persons in the top 10%, could be used to distinguish the person of interest from the remaining ones, by selecting new features, that are suitable for this task. Another possibility to handle a huge amount of people is to use cross camera probabilities and transition times, as in [11], to reduce the number of (probably similar) hypotheses.

5. Conclusion

We presented a novel approach for view invariant reidentification of persons by their appearances. It uses an automatic online feature selection mechanism based on the joint mutual information. Therefore, it is able to use only the most discriminative features for each person. During the search phase, the proposed method performs in hyper-real-time (10 seconds for 100 minutes video data). The experiments showed that our approach outperforms appearance-based state-of-the-art reidentification algorithms both regarding the reidentification quality and required computing time. We also showed, that score level fusion is easily applicable to our approach and improves the performance significantly.

References

[1] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *TNN*, 5(4):537–550, 1994. 4

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 2

[3] Casia gait database. http://www.sinobiometrics.com. obtained from http://www.cbsr.ia.ac.cn/english/gait. 5

[4] H. Cheng, X. Jiang, et al. Color image segmentation: advances and prospects. *PR*, 34:2259–2281, 2001. 3

[5] M. Farenzena, L. Bazzani, et al. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pp. 2360–2367, 2010. 2

[6] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *CVPR*, pp. 21–27, 1997. 3

[7] I. Guyon, S. Gunn, et al. *Feature Extraction: Foundations and Applications*. Springer, 2006. 4

[8] R. Haralick, K. Shanmugam, et al. Textural features for image classification. *TSMC*, 3:610–621, 1973. 3

[9] J. Hube. Methods for estimating biometric score level fusion. In *BTAS*, pp. 1–6, 2010. 5

[10] A. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, London, 1989. 3

[11] O. Javed, Z. Rasheed, et al. Tracking across multiple cameras with disjoint views. In *ICCV*, pp. 952–957, 2003. 7

[12] K. Juengling and M. Arens. Local feature based person reidentification in infrared image sequences. In *AVSS*, pp. 448–454, 2010. 5

[13] K. Juengling and M. Arens. View-invariant person re-identification with an implicit shape model. In *AVSS*, pp. 197–202, 2011. 2, 5, 6

[14] A. Leung and S. Gong. Online feature selection using mutual information for real-time multi-view object tracking. *AMFG*, 3723:184–197, 2005. 2

[15] D. Maltoni, D. Maio, et al. *Handbook of Fingerprint Recognition*. Springer, 2009. 5

[16] C. Martin, E. Schaffernicht, A. Scheidig, and H. Gross. Sensor fusion using a probabilistic aggregation scheme for people detection and tracking. In *ECMR*, pp. 176–181, 2005. 3

[17] Y. Ohta. *Knowledge-Based Interpretation of Outdoor Natural Color Scenes*. Pitman Publishing, London, 1985. 3

[18] T. Owens, K. Saenko, et al. Learning object color models from multi-view constraints. In *CVPR*, pp. 169–176, 2011. 7

[19] T. Pomierski and H. Gross. Biological neural architecture for chromatic adaptation resulting in constant color sensations. In *ICNN*, pp. 734 – 739, 1996. 3

[20] E. Schaffernicht, R. Kaltenhaeuser, S. Verma, and H. Gross. On estimating mutual information for feature selection. In *ICANN*, pp. 362–367, 2010. 3, 4

[21] K. Schenk, M. Eisenbach, A. Kolarow, K. Debes, and H. Gross. Comparison of laser-based person tracking at feet and upper-body height. In *KI*, pp. 277–288, 2011. 6

[22] K. Schenk, A. Kolarow, M. Eisenbach, K. Debes, and H. Gross. Automatic calibration of multiple stationary laser range finders using trajectories. In *AVSS*, 2012. 6

[23] D. Scott. *Multivariate density estimation: theory, practice and visualization*. John Wiley & Sons, New York, 1992. 4

[24] P. Shih and C. Liu. Comparative assessment of content-based face image retrieval in different color spaces. *AVBPA*, 3546:245–300, 2005. 3

[25] C. Stauffer and K. Tieu. Automated multi-camera planar tracking correspondence modeling. In *CVPR*, pp. I–259–266, 2003. 3

[26] S. Wang, M. Lewandowski, et al. Re-identification of pedestrians with variable occlusion and scale. In *ICCV Workshops*, pp. 1876–1882, 2011. 2

[27] J. Wu, C. Geyer, et al. Real-time human detection using contour cues. In *ICRA*, pp. 860–867, 2011. 2

[28] H. Yang and J. Moody. Feature selection based on joint mutual information. In *AIDA*, pp. 22–25, 1999. 2, 3