Person Re-Identification by Descriptive and Discriminative Classification

Martin Hirzer[1], Csaba Beleznai[2], Peter M. Roth[1], and Horst Bischof[1]

[1] Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{hirzer,pmroth,bischof}@icg.tugraz.at
[2] Austrian Institute of Technology, Austria
csaba.beleznai@ait.ac.at

Abstract. Person re-identification, i.e., recognizing a single person across spatially disjoint cameras, is an important task in visual surveillance. Existing approaches either try to find a suitable description of the appearance or learn a discriminative model. Since these different representational strategies capture a large extent of complementary information we propose to combine both approaches. First, given a specific query, we rank all samples according to a feature-based similarity, where appearance is modeled by a set of region covariance descriptors. Next, a discriminative model is learned using boosting for feature selection, which provides a more specific classifier. The proposed approach is demonstrated on two datasets, where we show that the combination of a generic descriptive statistical model and a discriminatively learned feature-based model attains considerably better results than the individual models alone. In addition, we give a comparison to the state-of-the-art on a publicly available benchmark dataset.

1 Introduction

Due to ceaseless advances in the research in semi-conductor, communications, and image sensors there is an increasing number of public areas that are subject to video surveillance. Thus, it becomes infeasible to analyze the ever growing amount of data – automatic systems are required. This especially applies for person re-identification, a central task in many surveillance scenarios, which can be described as recognizing an individual in different locations across a network of non-overlapping cameras. Besides of specific re-identification scenarios, e.g., tracking criminals over multiple cameras, typical tasks also include anonymous applications such as crowd analysis by identifying single instances. In general, this task has to be considered very challenging. Typical problems that have to be handled are extremely varying appearances of a person across the camera network (due to changing lighting conditions, different viewpoints, varying poses, etc.), people occluding each other, or a high number of very similar instances. Thus, motivated by the large number of practical applications and still unresolved problems there has been a considerable scientific interest within the last years.

For instance, Gheissari et al. [6] fit a triangulated graph to each individual to account for pose variations. However, the approach is only applicable for similar viewpoints. The same applies for the approach of Wang et al. [22], who segment an image of a

person into regions and capture their color spatial structure by a co-occurrence matrix. A more flexible approach was presented by Farenzena et al. in [4] exploiting perceptual principles relying on symmetry and asymmetry. They first run a segmentation step to obtain a person's silhouette and then accumulate the feature responses of color and texture features to a signature. Bird et al. [2] propose to segment the query image in equally spaced horizontal segments and extract the median HSL color of the foreground pixels of each of these segments.

In contrast, instead of designing specific features by hand, other methods aim to learn a suitable feature set or to directly generate a ranking model. Bak et al. [1] run a person detector and estimate a visual signature using Haar-like features that have been selected for each individual using AdaBoost. A similar but more sophisticated approach was presented by Gray and Tao [8]. They also select the most relevant features (color and texture) using AdaBoost but additionally estimate a likelihood ratio test for comparing corresponding features providing a similarity function. Lin et al. [13] and Schwartz et al. [18] propose to learn pairwise dissimilarities which can be applied for classification. Both approaches, however, require a training stage and labeled samples. Prosser et al. [16] formulate the person re-identification problem as a ranking problem. They introduce Ensemble RankSVM, which allows to learn a subspace where the potential true match gets the highest rank.

To further improve the classification results additional cues can be exploited. Makris et al. [14] and Rahimi et al. [17] simplify the problem by temporal reasoning on the spatial layout of the observed environment. Javed et al. [10] learn transitions between cameras to cope with problems such as illumination changes. Zheng et al. [23] enrich the description of persons by contextual visual information coming from the surrounding people.

These approaches can mainly be subdivided in two groups: (a) methods which employ a representation of descriptive statistics of the human appearance (using hand crafted features) [4, 6, 17, 22] and (b) approaches that are based on discriminative learning [3, 8, 13, 16, 18]. Thus, to take advantage of these complementary information cues, we apply the two strategies in parallel. First, we estimate a generic covariance-based description and calculate a similarity measure yielding a rank model. For examples that can not be classified in this way we compute a more specific discriminative model using boosting for feature selection. Moreover, we introduce a new covariance-based descriptor and adopt covariance features for the usage within a boosting framework. In the experimental results, we demonstrate the benefits of the proposed approach on two different datasets. In particular, we show that using a descriptive and discriminative model in parallel clearly improves the person re-identification capability. Additionally, we give a comparison to the state-of-the-art showing competitive results.

2 Person Re-Identification System

Given two camera views observing different locations of a scene, the goal of person re-identification is to select a certain person in one view and to recognize it in the other view. In the work on hand, we assume that we have already detected the persons in both views and we will refer the image of the selected person to as the *probe image* and

the images searched through the *gallery images* [7]. In particular, our system, which is illustrated in Figure 1, consists of a descriptive person model (see Section 3) and a discriminative person model (see Section 4), which are run consecutively.

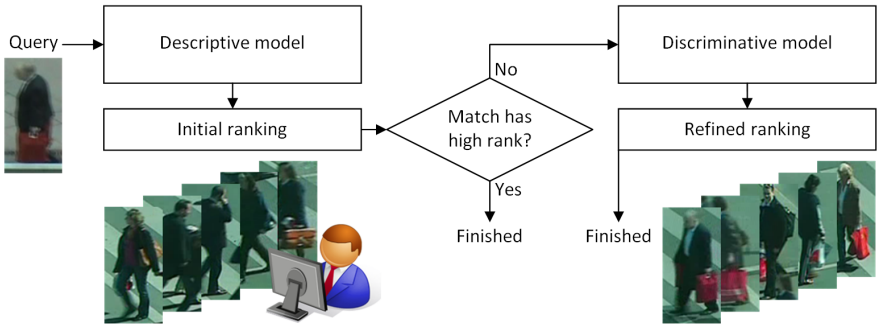

Fig. 1. Overview of the proposed system. After applying a descriptive model to obtain an initial ranking, a discriminative model can be used to refine the result.

For each probe image we first apply the descriptive person model to get an initial ranking of all gallery images. The first 50 images of this ranking are shown to a human operator, who then decides whether the searched person has been found or not. If not, we run the second stage, i.e., learn and evaluate the discriminative person model and rank the samples according to their confidence values. Since this model captures different aspects of an individual, focusing on details best separating it from others, there is a good chance that it can improve the ranking.

The descriptive model is based on a hand designed feature representation, hence, it can be estimated for any given single image. The discriminative model, however, is learned for each instance requiring positive and negative training data. Since we focus on person re-identification in a surveillance scenario, where multiple images of a person (multi-shot scenario) are available, we can use these images as training samples. If just one probe image is available (single-shot scenario), we can generate virtual samples using geometrical transformations and displacements. Hence, obtaining positive training samples is not much of a problem.

Though, for the negative training samples a more sophisticated sampling mechanism is required. For this purpose we use our descriptive model as starting point. As described before, applying this model already generates an initial ranked list of person images. Thus, we sample the negative images from the end of the list. Assuming that the descriptive person model provides a "good" ranking those images should be most dissimilar to the searched person. The overall principle is illustrated in Figure 2.

3 Descriptive Person Model

In the first stage of our person re-identification system we generate a descriptive statistical model which encodes visual appearance information. Considering the given task, the

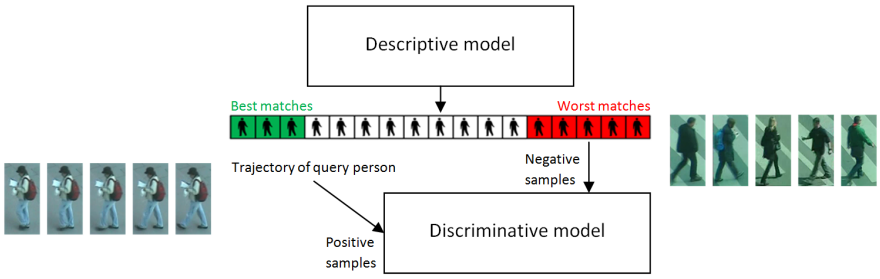

Fig. 2. Sampling of training images for the discriminative model. Positive samples are obtained from the trajectory of the query person, negative samples are drawn from the worst matches of the initial ranking provided by the descriptive model.

employed representation must meet requirements of specificity, invariance and computational efficiency. It implies that on the one hand the visual description must encompass discriminating visual information. On the other hand it must remain mostly unaffected in presence of photometric, view and pose changes. Moreover, for practical applicability the representation should be computed and matched rapidly at small memory requirements.

For our purpose we employ the region covariance descriptor of Tuzel et al. [20], which meets these criteria quite well. The descriptor is capable to combine multiple complementary cues, easy to compute and generates a compact signature. Since the descriptor aggregates several visual features, structural information of human visual appearance – such as the brightness relationship between upper and lower body halves – is represented only to a limited extent. In order to enhance the structural specificity of the representation, we use a set of covariance descriptors computed from multiple horizontal stripes covering the area of an image patch. This strategy is similar to the multiple region scheme used by [20] and to the principal axis histogram signature employed by [9].

For a given bounding box R with dimensions $W \times H$ a set of region covariance descriptors is computed in the following manner: The image within the bounding box $I_R(x, y)$ is used to compute a set of features, which represent intensity, color and texture. In order to capture spatial, color and gradient information, in our case the employed set of visual features comprises of

$$\{\mathbf{f}\} = \left[\mathbf{y}, \mathbf{L}, \mathbf{a}, \mathbf{b}, \left|\frac{\partial \mathbf{L}}{\partial \mathbf{x}}\right|, \left|\frac{\partial \mathbf{L}}{\partial \mathbf{y}}\right|\right], \tag{1}$$

i.e., the $\mathbf{y}$ pixel coordinate vector, the $\mathbf{L}$, $\mathbf{a}$, $\mathbf{b}$ color channels and the horizontal and vertical derivatives of the luminosity channel, respectively. The x-component of pixel coordinates is excluded from the feature set, thus allowing some invariance with respect to view variations when the person is seen from various sides.

The bounding box R is divided into N ($N = 7$ in our experiments) equally large horizontal stripes $\{S_l\}_{l=1..N}$ and within each stripe the covariance descriptor is computed

as

$$C^l = \frac{1}{z-1} \sum_{k=1}^{z} (\mathbf{f}_k - \mu)(\mathbf{f}_k - \mu)^T, \tag{2}$$

where C^l denotes the covariance matrix computed over z feature values within the l-th stripe and μ represents the vector of mean values computed on the individual features of the feature set.

The obtained set of covariance matrices $\{C^l\}_{l=1..N}$ defines a compact descriptor which encodes the interdependence between individual features computed inside the region of interest. A coarse structural information is captured using the set of covariances from multiple horizontal regions and by the weak spatial dependence given by the only slightly specific variation within the y-coordinate feature.

Similarity computation between two human appearances is performed by estimating the distance between two covariance matrices [5] in pairwise manner by

$$\rho\left(C_i^l, C_j^l\right) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k \left(C_i^l, C_j^l\right)}, \tag{3}$$

where C_i^l and C_j^l are computed for two different images i and j, but using the same stripe element with index l. λ_k denotes the generalized eigenvalues of C_i^l and C_j^l, and d is the number of features within the employed feature set (d = 6 in our case).

The covariance-based distance between two human appearances is defined as

$$\bar{d}_{ij} = \frac{1}{N} \sum_{l=1}^{N} \rho\left(C_i^l, C_j^l\right), \tag{4}$$

where $\bar{d}_{ij}$ is the mean covariance distance measure obtained from N stripe-versus-stripe comparisons. When a specific probe image is used as query, the probe image is compared to all gallery images and a set of distances is obtained. This set of distances is used to generate a ranking for every image in the gallery with respect to the probe.

4 Discriminative Person Model

In the second stage of our system we apply a discriminative model, which is estimated by Boosting for Feature Selection [19, 21]. Thus, similar to [1, 8] the goal is to select the most discriminant features for a specific instance from an over-complete feature set. However, unlike these methods, our approach does not involve any labeling of training data by hand. Moreover, the goal is not to learn a similarity function between image pairs but similar to [16] to finally generate a ranking of all gallery images. In particular, we train a model for each probe image and evaluate it on all gallery images. Those are then sorted according to their confidence values: a higher confidence results in a higher rank.

4.1 Estimating Ranks by Boosting for Feature Selection

Given a training set of positive and negative samples $\mathcal{S} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_L, y_L)\}$, where $\mathbf{x}_l \in \mathbb{R}^m$ is a sample and $y_l \in \{-1, +1\}$ is the corresponding label, a set of possible features $\mathcal{F} = \{f_1, ..., f_M\}$, a learning algorithm $\mathcal{L}$, and a weight distribution D, that is initialized uniformly by $D(l) = \frac{1}{L}$. Then, the main idea of boosting for feature selection is that each feature f_j corresponds to a single weak classifier h_j and that boosting selects an informative subset of N features. In each iteration n, $n = 1, ..., N$ all features f_j, $j = 1, \ldots, M$ are evaluated on all samples $(\mathbf{x}_l, y_l)$, $l = 1, \ldots, L$ and hypotheses are generated by applying the learning algorithm $\mathcal{L}$ with respect to the weight distribution D over the training samples. The best hypothesis is selected and forms the weak classifier h_n. The weight distribution D is updated according to the error of the selected weak classifier.

The process is repeated until N features are selected, i.e., N weak classifiers are trained (N = 20 in our experiments). Finally, we estimate a confidence measure[3] C according to a weighted linear combination of all weak classifiers h_n:

$$C(x) = \sum_{n=1}^{N} \alpha_n h_n(x). \tag{5}$$

4.2 Features

Due to the popularity various different features, e.g., Haar-like [21], Edge Orientation Histograms [12], or boundary fragments [15] have been introduced for the application with boosting for feature selection. Such features mainly capture generic visual object properties and have shown excellent performance for object recognition/detection and tracking. However, for the re-identification task they are often not discriminative enough. In particular, as also discussed in Section 5.1, we found that the most important information queues are intensity changes between the upper and lower body of a person and color. Thus, for our application we use a combination of horizontally divided Haar features and covariance features. Moreover, to avoid that too much background information is modeled by the (local) features we prohibit features that are placed close to the image borders.

Since Haar features are well known in the context of boosting (e.g., [21]) in the following we focus on the discussion on the covariance features capturing the essential color information. As described in Section 3, covariance matrices, in general, provide an elegant way of integrating various different feature channels, in our case RGB color channels, into one compact representation. They capture the variance of these channels and the correlation between them. However, since the space of covariance matrices does not form a Euclidean space they cannot directly be used in a boosting framework. To overcome this limitation, we follow the approach described in [11], allowing to describe the covariance matrices in a Euclidean vector space. In particular, for the d-dimensional case a set of $2d + 1$ specific vectors $\mathbf{s}_i \in \mathbb{R}^d$, called *Sigma Points*, is constructed as follows:

[3] If required a strong classifier H can be estimated by $H(x) = sign\,(C(x))$.

$$\mathbf{s}_0 = \mu \qquad \mathbf{s}_i = \mu + \alpha(\sqrt{C})_i \qquad \mathbf{s}_{i+d} = \mu - \alpha(\sqrt{C})_i, \tag{6}$$

with $i = 1 \ldots d$, μ and C being the data's mean vector and covariance matrix respectively, and $(\sqrt{C})_i$ being the i-th column of the covariance matrix square root. The scalar α is a constant weighting for the elements in the covariance matrix and is set to $\alpha = \sqrt{2}$ for Gaussian data. The points $\mathbf{s}_i$ accurately capture the statistics of the original covariance matrix up to third order for Gaussian and up to second order for non-Gaussian data. The final feature representation is built by concatenation of all *Sigma Points* into one vector. Hence, *Sigma Points* provide a very powerful representation that is capable of integrating various different feature channels into one compact feature vector.

With this representation we are now able to efficiently capture local color information in our boosting algorithm. As for Haar features, we use a rectangular shaped region for extracting color information (RGB) from an image. All pixels within the covariance feature's region are used to calculate the mean vector μ, covariance matrix C and finally the *Sigma Points* representation. This enables us to capture very discriminative, local color features of a person (e.g., red bag), as opposed to the descriptive statistical model described in Section 3, which extracts color and gradient information from regular stripe regions laid over the person image. As weak learner h_j we apply a Bayesian decision criterion for the Haar features and a multidimensional nearest neighbor classifier for the *Sigma Points*. Haar and covariance features are illustrated in Figure 3.

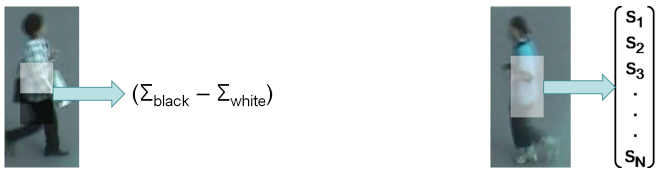

Fig. 3. Applied features: (a) Haar features mainly capture intensity changes between the upper and lower body of a person, (b) covariance features extract local color information in form of vectors of *Sigma Points*.

5 Experimental Results

We evaluated our approach on two datasets[4], the public VIPeR dataset [7] (single-shot scenario) and our own person re-identification dataset[5] (multi-shot scenario). Examples of both are shown in Figure 4. As performance measure we use Cumulative Matching Characteristic (CMC) curves [22], which represent the expectation of the true match being found within the first n ranks.

[4] Other benchmarks have been proposed (e.g., [1,4,16]), however, since either no annotations are available or the datasets are not uniquely defined, we did not used them for our experiments.

[5] Available at http://lrs.icg.tugraz.at/downloads.php.

(a) (b)

Fig. 4. Example image pairs from the VIPeR dataset (a) and example trajectory images from our multi-frame dataset (b). Upper and lower row correspond to different camera views.

5.1 VIPeR Dataset

The VIPeR dataset consists of 632 person image pairs taken from two different camera views. Most of the example pairs contain a viewpoint change of about 90 degrees as well as significant changes in pose and illumination, making person re-identification very challenging. To compare our method to other approaches, we followed the evaluation procedure described in [4, 8]. The authors split the set of 632 image pairs randomly into two sets of 316 image pairs each, one for training and one for testing, and build the average over several runs. Since we do not need a training set, we evaluate our algorithm on a subset of 316 randomly selected image pairs and also average the results of several runs. Considering images from one camera as the probe set, and images from the other camera as the gallery set, we match each probe image with all images from the gallery set.

When applying our discriminative person model we need positive and negative training samples for the boosting step. In our scenario positive training samples are extracted from person trajectories, and negative training samples are drawn from the gallery images that received the lowest ranks in the initial, descriptive ranking step. However, the VIPeR dataset does not provide trajectories, just image pairs. Thus, we generate virtual positive training images from the probe image by randomly applying slight geometric distortions and smoothing. Figure 5 and Table 1 show the average results of our approach on the VIPeR dataset of 5 runs on randomly selected subsets of 316 image pairs.

As one can see, the descriptive and discriminative person model have similar performance. However, since they describe different aspects of a person, taking into account both models yields a significant improvement. This is shown by a third curve that is generated using the model returning the higher match rank for each probe image, simulating the human operator decision described in Section 2. Moreover, in Table 1 we compare the performance of our approach on the range relevant for our approach, i.e., the first 50 ranks, to state-of-the-art methods [4, 8, 16]. As can be seen we obtain competitive results, especially, for rank 1. Even though in contrast to [8, 16] we do not need any (hand) labeling of data and unlike [4] we do not use a foreground-background segmentation. However, we expect that using a segmentation step will improve our results notably, especially in cases of great pose variations, e.g., varying leg postures.

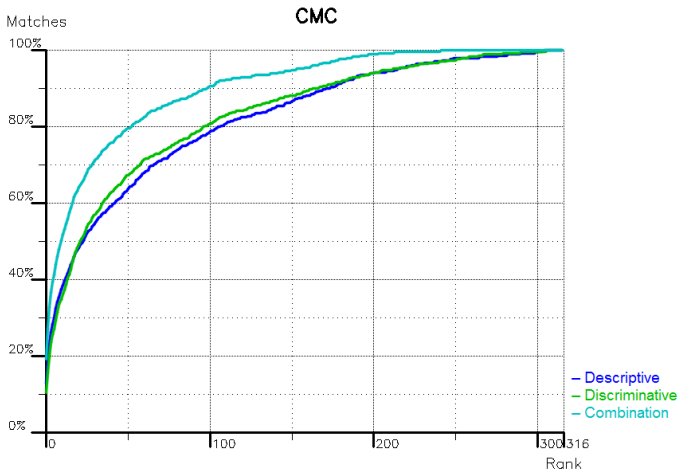

Fig. 5. CMC curves of our approach on the VIPeR dataset. The blue curve shows the descriptive and the green curve shows the discriminative person model. The combination of both models is depicted in cyan color.

Rank	ELF	SDALF	ERSVM	Our Approach
1	12%	20%	13%	19%
10	43%	50%	50%	52%
25	66%	70%	71%	69%
50	81%	85%	85%	80%

Table 1. Matching rates for ELF, SDALF, ERSVM and our algorithm on the VIPeR dataset.

As discussed in Section 4, we apply only Haar and covariance features for our re-identification task. In the following, we illustrate that exactly these features are best suited for our task by evaluating different features on the first 30 image pairs of the VIPeR dataset: Haar-like, histograms of oriented gradients (HOGs), local binary patterns (LBPs), covariance features using RGB channels, as well as their combinations. The obtained results in form of CMC curves are depicted in Figure 6. It can clearly be seen that color (captured by covariance features) is the strongest cue, followed by Haar-like features, which particularly capture intensity changes between the upper and lower body of a person. HOGs and LBPs, on the other hand, perform rather poorly, since they concentrate on finer structures that are often not visible in the gallery image due to viewpoint changes. In fact, the best performance was achieved using a combination of Haar-like and covariance features.

5.2 Multi-Shot Dataset

Since the intended use case for the proposed method was to apply person re-identification on surveillance data, we generated a multi-shot dataset. It consists of images extracted

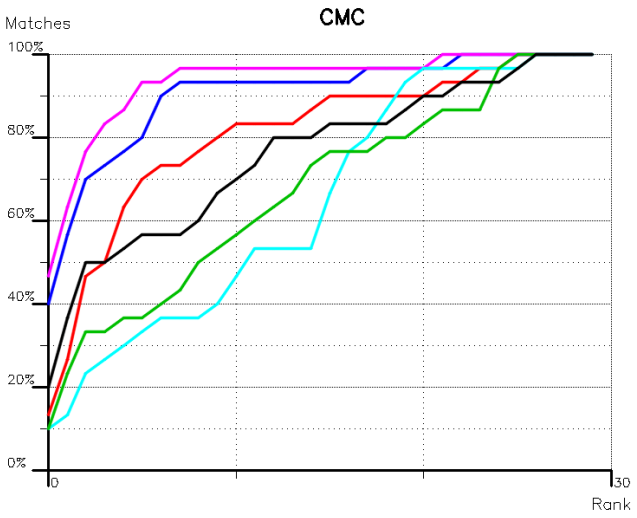

Fig. 6. Different feature types evaluated on the first 30 image pairs of the VIPeR dataset: Haar (red), HOG (green), LBP (cyan), Covariance (blue), Haar + Covariance (magenta), combination of all types (black)

from multiple person trajectories recorded from two different static surveillance cameras. Images from these cameras contain a viewpoint change and a stark difference in illumination, background and camera characteristics (e.g., green cast). Since images are extracted from trajectories, several different poses per person are available in each camera. We have recorded 475 person trajectories from one camera and 753 from the other one, with 245 persons appearing in both views. Thus, each of the 245 persons in the probe set is searched in a gallery set of 753 individuals. Each trajectory consists of approximately 100 to 150 images, depending on the walking speed of an individual. For the gallery set we equidistantly extracted 5 images per trajectory. The maximum rank returned by these 5 images defines the rank of a person.

On this dataset, positive samples for the boosting step can easily be extracted from the trajectory of the searched person. To get some additional variation into the positive training set, we also generate a few virtual samples, as for the VIPeR dataset. To acquire negative training samples we again use the ranked list of gallery images provided by our descriptive model. For the features used in the boosting step we use the same setup as for the VIPeR dataset.

Figure 7 shows the average results of our approach on this dataset after 3 runs. As shown by the curves, in contrast to the VIPeR image pairs, the discriminative model slightly outperforms the descriptive model. This can be explained by greater variability captured if positive training samples are extracted from a whole trajectory. Thus, an overfitting to the small number of positive samples can be prevented. Finally, like on the VIPeR dataset, taking into account descriptive and discriminative information leads to superior performance.

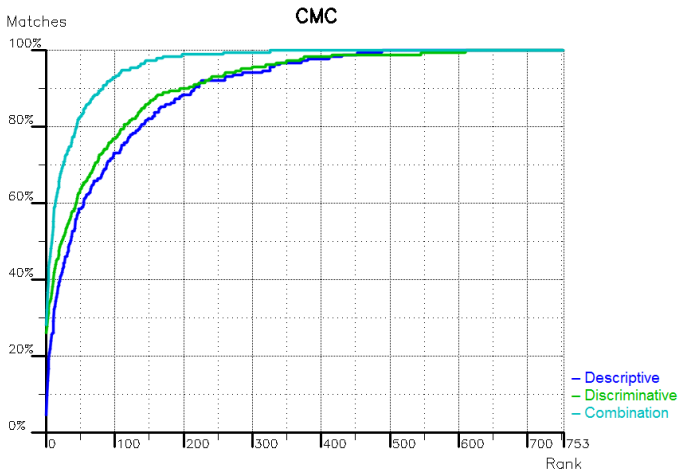

Fig. 7. CMC curves of the proposed algorithm on our multi-frame dataset. The blue curve shows the descriptive and the green curve shows the discriminative person model. The combination of both models is depicted in cyan color.

6 Conclusion

Typical approaches for person re-identification either estimate a visual signature describing the appearance of a query sample or train a discriminative model. In this paper we took advantage of both approaches and introduced a system combining descriptive and discriminative models. We first run an appearance-based matcher using a covariance description, which has shown to be a considerable trade-off between speed and accuracy. For examples where this representation exhibits low specificity in a second stage a discriminative model is estimated by boosting for feature selection. In particular, we found that two types of features describing intensity transitions and color information (i.e., Haar features and *Sigma Points*) are best suited for the given task. The experimental results demonstrated that compared to the single cues using the proposed approach significatly better results can be obtained. In addition, we gave a comparison to state-of-the-art methods on a publicly available dataset. Even though avoiding any labeling and having only a limited amount of training data we can report competitive results.

Acknowledgments. This work has been supported by Siemens AG Österreich, Corporate Technology (CT T CEE), Austria, and the project SECRET (821690) under the Austrian Security Research Programme KIRAS.

References

1. S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-idendification using Haar-based and DCD-based signature. In *Workshop on Activity Monitoring by Multi-Camera Surveil-*

lance Systems, 2010.
2. N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Trans. Intelligent Transportation Systems*, 6(2):167–177, 2005.
3. O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010.
4. M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. CVPR*, 2010.
5. W. Förstner and B. Moonen. A metric for covariance matrices. Technical report, Department of Geodesy and Geoinformatics, Stuttgart University, 1999.
6. N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. CVPR*, 2006.
7. D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. PETS*, 2007.
8. D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. ECCV*, 2008.
9. M. Hu, J. Lou, W. Hu, and T. Tan. Multicamera correspondence based on principal axis of human body. In *Proc. ICIP*, 2004.
10. O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. CVPR*, 2005.
11. S. Kluckner, T. Mauthner, P. M. Roth, and H. Bischof. Semantic classification in aerial imagery by integrating appearance and height information. In *Proc. ACCV*, 2009.
12. K. Levi and Y. Weiss. Learning object detection from a small number of examples: The importance of good features. In *Proc. CVPR*, 2004.
13. Z. Lin and L. S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *Advances Int'l Visual Computing Symposium*, 2008.
14. D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Proc. CVPR*, 2004.
15. A. Opelt and A. Z. Axel Pinz. A boundary-fragment-model for object detection. In *Proc. ECCV*, 2006.
16. B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. BMVC*, 2010.
17. A. Rahimi, B. Dunagan, and T. Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *Proc. CVPR*, 2004.
18. W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Proc. Brazilian Symposium on Computer Graphics and Image Processing*, 2009.
19. K. Tieu and P. Viola. Boosting image retrieval. In *Proc. CVPR*, 2000.
20. O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. ECCV*, 2006.
21. P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, 2001.
22. X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *Proc. ICCV*, 2007.
23. W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proc. BMVC*, 2009.