Person re-identification by pose priors

Sławomir Bąk Filipe Martins Francois Brémond

INRIA Sophia Antipolis, STARS team, 2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex - France
firstname.surname@inria.fr

ABSTRACT

The person re-identification problem is a well known retrieval task that requires finding a person of interest in a network of cameras. In a real-world scenario, state of the art algorithms are likely to fail due to serious perspective and pose changes as well as variations in lighting conditions across the camera network. The most effective approaches try to cope with all these changes by applying metric learning tools to find a transfer function between a camera pair. Unfortunately, this transfer function is usually dependent on the camera pair and requires labeled training data for each camera. This might be unattainable in a large camera network. In this paper, instead of learning the transfer function that addresses all appearance changes, we propose to learn a generic metric pool that only focuses on pose changes. This pool consists of metrics, each one learned to match a specific pair of poses. Automatically estimated poses determine the proper metric, thus improving matching. We show that metrics learned using a single camera improve the matching across the whole camera network, providing a scalable solution. We validated our approach on a publicly available dataset demonstrating increase in the re-identification performance.

Keywords: re-identification, metric learning, pose matching

1. INTRODUCTION

Person re-identification is a well known problem in computer vision community. This task requires finding a target appearance in a network of cameras with non-overlapping fields of view. The changes in person pose together with different camera viewpoints and different color responses make the task of appearance matching extremely difficult.

Current state of the art approaches focus either on *feature modeling*[1–4] that designs descriptors invariant to camera changes or on *metric learning*[5–11] that uses training data to search for matching strategies minimizing the appearance changes (intra-class variations), while highlighting distinctive properties of the target (maximizing inter-class variation). Feature-modeling approaches concentrate on feature representation which should be invariant to pose and camera changes. These approaches usually assume *a priori* an appearance model, focusing on designing novel features for capturing the most distinctive aspects of an individual. Nevertheless, metric learning approaches are often the one that achieve the best performance in re-identifying people. These approaches learn a distance function that transfer the feature space from one camera to the other such that relevant dimensions are emphasized while irrelevant ones are discarded. Although, this transfer function boosts the recognition accuracy, it is usually camera pair dependent and requires large training data (hundreds of annotated image pairs with the same individual) for each camera pair. This might be unattainable in a large camera network. Moreover, metric learning can lose the performance while directly computing the difference between two feature vectors without aligning them first.

In this paper, instead of learning a single metric that tackles all difficulties related to appearance changes, we propose to divide the person re-identification task into several sub-tasks: (1) target alignment; (2) pose estimation and (3) pose-driven **metric pool**.

The first two sub-tasks: target alignment and pose estimation are approached using the recently published method[12] (see Section 2.1). From Fig. 1 we can notice that serious perspective changes might significantly distort the appearance, deteriorating the recognition accuracy. Fortunately, a simple but efficient affine transformation using 3D scene information can improve alignment of two images extracted from different cameras. The pose orientation can be automatically estimated using the motion or the trajectory of the target.

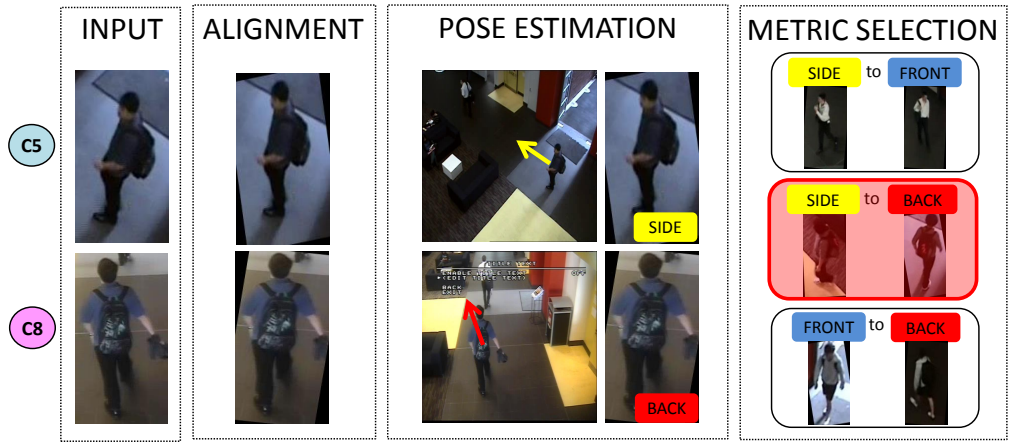


Figure 1. **Person re-identification using pose priors**: the figure illustrates the process of matching the same subject across different camera views; the cropped images are first aligned, then the pose orientations are estimated determining the metric for the given pose change. The selected metric reflects the specific to pose change transformation of the feature space.

Given two images with estimated poses of the subject, it is highly desirable to develop the strategy that could exploit pose information for improving the matching accuracy. In the result we define the third task that consists in generating a pool of metrics, each one learned to match a specific pair of poses. In this paper we propose to learn the metric pool using only a single camera. We believe that this avoids over-fitting the metric to the given camera pair. Once the metric pool is learned, it can be applied to any pair of cameras. While matching two images, we select the proper metric based on automatically estimated pose of the target image and of the record image in the gallery (database). The selected metric reflects the transformation of the feature space between two given poses, thus improving matching. The proposed solution showed to be effective and scalable (see Sec. 3).

2. THE METHOD

Our approach requires two steps. The first stage overcomes perspective changes and estimates the pose of the target. We apply a simple but effective technique[12] that minimizes perspective distortions by an affine transformation of a cropped image containing the target. The pose is estimated using 3D scene information and motion of the target. The second stage is responsible for appearance matching using a metric pool that consists of metrics, each one learned to match a specific pair of poses.

2.1 Target alignment and pose estimation

The changing viewpoint in a network of cameras might significantly distort the visual appearance of a person. We minimize perspective distortions by rotating the cropped image with the person by an angle computed by mapping the ground plane living in real world coordinates to the vertical of a given image (see Ref. 12). This mapping can be computed by employing the calibration information of the camera (*e.g.* Tsai calibration). In the result, the cropped image should contain the person aligned to the vertical axis (see ALIGNMENT in Fig. 1).

We estimate the pose by using 3D scene information and motion of the target.[12] Pose orientation is computed as a dot product between the viewpoint vector and the motion vector. The view point vector is defined as the difference between the real world location coordinates of the target in the scene and the real world location coordinates of the camera, both projected on the ground plane. The motion vector is determined by the difference between two consecutive real world location coordinates of the target. The dot product stands for the angle between the camera viewpoint and the pose direction of the target. Employing this simple but effective method, we obtain sufficiently accurate information for estimating the pose. This information is used to select the data for training a metric pool for pose changes and then it can be used to select the proper metric while matching different poses.

2.2 Metric pool

Given two images with estimated poses of the subject, we develop the strategy that can exploit pose information, thus improving the matching. This strategy consists in generating a pool of metrics, each one learned to match a specific pair of poses. For learning metrics we employ Mahalanobis metrics, which have recently attracted a lot of research interest in computer vision.

2.2.1 Mahalanobis metric

As state of the art we can consider KISS metric,[5] Large Margin Nearest Neighbor Learning (LMNN),[8] Information Theoretic Metric Learning (ITML)[6] and Logistic Discriminant Metric Learning (LDML).[7] These machine learning algorithms learn a distance or similarity metric based on the class of Mahalanobis distance functions. Mahalanobis-like metric measures the squared distance between two data points x_i and x_j:

$$d^2_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j), \tag{1}$$

where $\mathbf{M} \geq 0$ is a positive semi-definite matrix and $x_i, x_j \in \mathbb{R}^d$ is a pair of samples (i, j). Metric learning methods rely on the idea that similar data points should be closer in the feature space than dissimilar points. For training we need a similarity label $y_{ij} : y_{ij} = 1$ for similar pairs, *i.e.* if the samples share the same class label $(y_i = y_j)$ and $y_{ij} = 0$ otherwise.

Usually metric learning approaches rely on an iterative optimization scheme which can get computationally expensive for large datasets. In contrast to these methods, KISS metric is a non-iterative method that builds on a statistical inference perspective of the space of pairwise differences.[5] In the result, it is orders of magnitudes faster than comparable metric learning approaches.

2.2.2 Learning a pose-change metric

Let us assume that the pose can be described by the orientation angle between the motion vector of the target and the viewpoint vector of the camera (see details in Sec. 2.1). Thus, for each image we have given the pose as the angle in the range of $[0°, 360°)$. We divide this range into n bins, *i.e.* pose $p \in P$, where $|P| = n$. Given n bins of estimated poses, the idea is to learn metrics that stand for transfer functions between pairs of two poses. In the result, the metric pool will consist of $\frac{n(n+1)}{2}$ metrics, each one learned to match specific pair of poses.

Learning is performed in the following way. For each pair of poses, we automatically select subjects that support a given pose transition taking place in the same camera. We learn a pose-change metric using only a single camera to avoid including any color transfer in the metric. While learning metrics, we follow a well known scheme based on image pairs, containing two desired poses of the same target. Let us assume that we want to learn the metric for the pose change from pose a to pose b. In this case $y_{ij} = 1$ only if it is the same subject $(y_i = y_j)$ and only if it supports the pose change $(p_i = a \wedge p_j = b)$, $y_{ij} = 0$ otherwise.

For learning metrics we employ the previously mentioned metric learning tools.[5–8] As the learning is performed offline, the time complexity is not the main concern. However, if the reader is interested in training on a large dataset we recommend the KISS metric.[5] In experiments we compare the performance of all these approaches.

The set of learned metrics stands for the metric pool. This metric pool is not dependent on a camera pair. Thus, once the pool for matching poses is learned, it can be applied to any pair of cameras.

While matching two images given from different (or the same) camera, we first align subjects and estimate their pose orientations. Having two poses, we select the corresponding metric from the metric pool. This metric is used to compute the similarity between given subjects that is used in the final ranking. As the selected metric reflects the transformation of the feature space between two given poses, it improves the recognition accuracy.

3. EXPERIMENTAL RESULTS

This section evaluates the re-identification performance, while employing metric learning tools for handling pose changes. The results are analyzed in terms of recognition rate, using the cumulative matching characteristic (CMC)[13] curve. The CMC curve represents the expectation of finding the correct match in the top n matches.

3.1 Database and experimental setup

We carry out experiments on the SAIVT-SOFTBIO database.[14] This dataset consists of 152 people moving through a network of 8 cameras. Subjects travel in an uncontrolled manner, thus most of subjects appear only in a subset of the camera network. This provides a highly unconstrained environment reflecting a real-world scenario. In average, each subject is registered by 400 frames spanning up to 8 camera views in challenging surveillance conditions (significant illumination, pose and viewpoint changes). Provided annotations given by coarse bounding boxes indicate locations of the subjects in each frame. The centers of these bounding boxes build trajectories of the subjects. Thanks to trajectories and 3D scene information we align and estimate the pose of subject images.[12]

3.1.1 Appearance model

Every cropped and aligned image is scaled into a fixed size window of 64×192 pixels. A set of rectangular sub-regions is produced by shifting 32×32 regions with a 16 pixels step. This operation results in 33 overlapping rectangular sub-regions. From each sub-region, we extract RGB color and HOG histograms. We minimize color dissimilarities caused by camera illumination changes by applying *histogram equalization* to each color channel. By this operation we try to avoid a dependency of our metric on the camera color spectrum.

3.1.2 Learning and testing

For learning a metric we selected the camera 5 as it provides the sufficient number of subjects (37) and images (279) that support a given pose change. The pose orientation was divided into $n = 3$ bins: the front pose, the back pose and the side pose (either left or right as the image can be flipped). The centers of bins are $0°, 180°$ and $90°$, respectively. The image is classified in one of the poses based on the nearest neighbor strategy.

We learn the transfer from the side pose to the back pose using a single camera. By using a single camera we want to avoid including any color transfer in our metric, thus producing independent to camera pair metric. Training data was processed using the method Ref. 12 that crops, aligns and estimates pose of the subject in each frame.

Learning is performed using KISSME framework[5] that provides several metric learning tools (KISS metric,[5] Mahalanobis distance with similar pairs, Large Margin Nearest Neighbor Learning (LMNN),[8] Information Theoretic Metric Learning (ITML)[6] and Logistic Discriminant Metric Learning (LDML).[7]

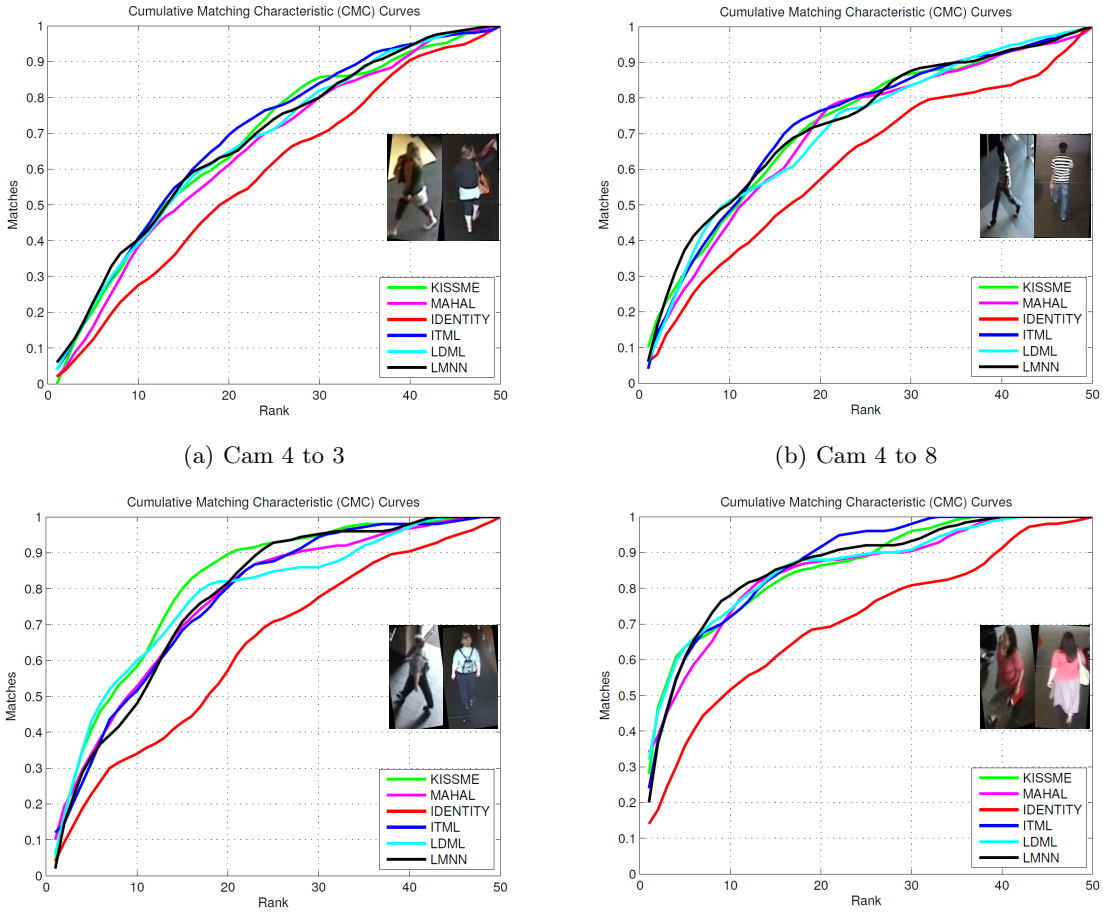
For testing we randomly select a single image for each subject that supports the given pose change. All camera pairs are evaluated using 50 subjects. The procedure is repeated 10 times to obtain reliable statistics.

3.2 Results

Fig. 2 illustrates the results of re-identification on different camera pairs. IDENTITY label corresponds to the diagonal metric $\mathbf{M}$ that is the Euclidean distance (L_2 metric). The remaining labels correspond to state of the art metric learning tools: KISSME,[5] MAHAL-Mahalanobis for $y_{ij} = 1$, ITML,[6] LDML[7] and LMNN[8]).

From the results, it is apparent that learning a metric for the pose change improves the recognition for all camera pairs. Each metric learning method shows improvement *w.r.t* the L_2 metric. This result is very promising, especially in Fig. 2(a,b) where we can notice the improvement even when the camera pair does not contain the training camera 5.

From Fig. 2(c,d) it seems that the improvement in matching is higher when the testing camera pair contains the camera 5. This might mean that our metric show some dependency on the training data (*e.g.* the selected model is dependent on the features available in the camera 5). In the result, while learning a metric pool we should properly select the training data to obtain sufficiently general metrics. Alternative solution would be to choose such image features for representing the appearance model that are less dependent on the camera. We performed



(a) Cam 4 to 3

(b) Cam 4 to 8

(c) Cam 5 to 3

(d) Cam 5 to 8

Figure 2. **Person re-identification by pose change metric**: CMC curves obtained on the SAIVT-SOFTBIO dataset on different camera pairs. The metric for the change from the side pose to the back pose was learned using images from the camera 5.

tests using only the HOG descriptor, but unfortunately we obtained decrease in the performance. Designing a new appearance model, we need to look for a trade-off between its discriminative power and invariance through cameras. This task is particularly hard, especially, as this trade-off varies from data to data.[15]

4. CONCLUSION AND PERSPECTIVES

We proposed to re-identify people using their pose orientations. Our strategy is based on the idea that the appearances of the subject in different poses ought to be matched by different metrics. The pose can be estimated using the motion of the target. Given estimated poses, we learned the metric pool, *i.e* a set of metrics where each one is learned to match a specific pair of poses. We learned the transfer functions employing Mahalanobis metrics using only a single camera. This allowed us to apply the metric to uncorrelated camera pairs, providing the scalable solution for large camera networks. Experiments on various cameras demonstrated that our method consistently improves the re-identification accuracy. In the future, we will further explore the generalization capability of the pose-driven metric pool. Different training schemes will be tested to obtain more general metrics. Additionally, we plan to analyze the number of metrics in the metric pool *w.r.t.* the re-identification accuracy and propose a finer mapping from the image to the pose using a depth sensor.

ACKNOWLEDGMENTS

This work has been supported by PANORAMA, CENTAUR and MOVEMENT European projects.

REFERENCES

[1] Bak, S., Corvee, E., Bremond, F., and Thonnat, M., "Person re-identification using spatial covariance regions of human body parts," in [*AVSS*], (2010).

[2] Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M., "Person re-identification by symmetry-driven accumulation of local features," in [*CVPR*], (2010).

[3] Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P., "Shape and appearance context modeling," in [*ICCV*], (2007).

[4] Bazzani, L., Cristani, M., Perina, A., Farenzena, M., and Murino, V., "Multiple-shot person re-identification by hpe signature," in [*ICPR*], (2010).

[5] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H., "Large scale metric learning from equivalence constraints," in [*CVPR*], (2012).

[6] Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S., "Information-theoretic metric learning," in [*ICML*], (2007).

[7] Guillaumin, M., Verbeek, J., and Schmid, C., "Is that you? metric learning approaches for face identification," in [*ICCV*], (2009).

[8] Weinberger, K. Q., Blitzer, J., and Saul, L. K., "Distance metric learning for large margin nearest neighbor classification," in [*NIPS*], (2006).

[9] Dikmen, M., Akbas, E., Huang, T. S., and Ahuja, N., "Pedestrian recognition with a learned metric," in [*ACCV*], (2010).

[10] Zheng, W.-S., Gong, S., and Xiang, T., "Person re-identification by probabilistic relative distance comparison," in [*CVPR*], (2011).

[11] Li, W. and Wang, X., "Locally aligned feature transforms across views," in [*CVPR*], (2013).

[12] Bak, S., Zaidenberg, S., Boulay, B., and Bremond, F., "Improving Person Re-identification by Viewpoint Cues," in [*AVSS*], (2014).

[13] Gray, D., Brennan, S., and Tao, H., "Evaluating Appearance Models for Recognition, Reacquisition, and Tracking," in [*PETS*], (2007).

[14] Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., and Lucey, P., "A database for person re-identification in multi-camera surveillance networks," in [*DICTA*], (2012).

[15] Varma, M. and Ray, D., "Learning the discriminative power-invariance trade-off," in [*ICCV*], (2007).