Local Descriptors Encoded by Fisher Vectors for Person Re-identification

Bingpeng Ma, Yu Su, Frédéric Jurie

▶ **To cite this version:**

Bingpeng Ma, Yu Su, Frédéric Jurie. Local Descriptors Encoded by Fisher Vectors for Person Re-identification. 12th European Conference on Computer Vision (ECCV) Workshops, 2012, Italy. pp.413-422, 10.1007/978-3-642-33863-2_41 . hal-00806066

HAL Id: hal-00806066
https://hal.science/hal-00806066

Submitted on 29 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Local Descriptors encoded by Fisher Vectors for Person Re-identification

Bingpeng Ma, Yu Su, and Frédéric Jurie

GREYC — CNRS UMR 6072, University of Caen Basse-Normandie, Caen, France
{bingpeng.ma,yu.su,frederic.jurie}@unicaen.fr

Abstract. *This paper proposes a new descriptor for* person re-identification *building on the recent advances of Fisher Vectors. Specifically, a simple vector of attributes consisting in the pixel coordinates, its intensity as well as the first and second-order derivatives is computed for each pixel of the image. These local descriptors are turned into Fisher Vectors before being pooled to produce a global representation of the image. The so-obtained Local Descriptors encoded by Fisher Vector (LDFV) have been validated through experiments on two person re-identification benchmarks (VIPeR and ETHZ), achieving state-of-the-art performance on both datasets.*

1 Introduction and related works

In recent years, person re-identification in unconstrained conditions have attracted more and more research interest. Person re-identification consists in recognizing an individual through different images (e.g. coming from cameras in a distributed network or from the same camera at different time). The key issue of such systems is their ability to measure the similarity between two person-centered bounding boxes and predict if they represent to the same person, despite changes in illumination, viewpoint, background, occlusions and low resolution.

In order to tackle this problem, many researchers have concentrated on (i) the design of visual features to describe individual images, and (ii) the use of adapted distance measures (e.g. obtained by metric learning), to predict if two images represent the same person.

The visual features applied in person re-identification can be roughly categorized into global and local features. While global features encode the holistic configuration of body parts, local features encode the detailed traits within body parts. The typical features for person re-identification are: color (widely used since the color of clothing constitutes simple but efficient visual signatures) [1], HOG like signatures [2, 3], texture [4–6], differential filters [6], Haar-like filters [7], co-occurrence matrices [3], interest points [8], e.g. SURF and SIFT [9, 10] or the signatures of image regions [2, 1]. In addition, these features can be combined as they play different roles. For example, [4] combined 8 color features with 21 texture filters (Gabor and differential filters). [1] and [11] combined Maximally Stable Color Regions (MSCR) descriptors with weighted color

histograms, achieving the state-of-the-art results on several wildly-used person re-identification datasets under unsupervised setting.

Perhaps the most common approach for combining local features for producing a global signature of the image is the Bag-of-Words (BoW) model [12], in which local features extracted from an image are first mapped to a set of visual words and then the image is represented as a histogram of visual word occurrences. The BoW model has been used for person re-identification in [10], where the authors builds groups of descriptors by embedding the visual words into concentric spatial structures and by enriching the BoW description of a person by the contextual information coming from the surrounding people. Recently, the BoW model has been greatly enhanced by the introduction of the Fisher vector [13] which encodes higher order statistics of local features. It has been shown that the resultant Fisher vector gives excellent performance for several challenging object recognition and image retrieval tasks [14, 15].

In addition, metric learning can be used to further improve the performance by providing a metric adapted to the task (e.g. [16, 10, 4]). Most distance metrics learning approaches learn a Mahalanobis-like distance, such as Large Margin Nearest Neighbors (LMNN) [17], Information Theoretic Metric Learning (ITML) [18] and Logistic Discriminant Metric Learning (LDML) [19], and Pairwise Constrained Component Analysis (PCCA) [20].

Building on these advances, this paper proposes to combine Fisher vectors with a new very simple 7-d local descriptor adapted to the representation of persons images, and to use the resultant representation (*Local Descriptors encoded by Fisher Vector* or LDFV) to describe persons images. Specifically, in LDFV, each pixel of an image is converted into a 7-d local feature, which contains the coordinates, the intensity, the first-order and second-order derivative of this pixel. Then, the local features are encoded and aggregated into a global Fisher vector, *i.e.* the LDFV representation. Finally, we learn the distance between LDFV representations using the metric learning approach proposed by [20].

The proposed representation has been experimentally validated on two person re-identification databases (VIPeR and ETHZ), which are challenging since they contain pose changes, viewpoint and lighting variations, and occlusions. Furthermore, these datasets have been used in the recent literature, allowing comparisons with recent approaches.

2 Description of the proposed approach

2.1 LDFV: Local Descriptor encoded by Fisher Vector

In order to capture the local properties of images, we have designed a very simple 7-d descriptor inspired by [21]:

$$f(x, y, I) = (x, y, I(x, y), I_x(x, y), I_y(x, y), I_{xx}(x, y), I_{yy}(x, y)) \quad (1)$$

where x and y are the pixel coordinates, $I(x, y)$ is the raw pixel intensity at position (x, y), I_x and I_y are the first-order derivatives of image I with respect to x and y, while I_{xx} and I_{yy} are the second-order derivatives.

Let $M = \{m_t, t = 1, \ldots, T\}$ be the set of the T local descriptors extracted from an image. The key idea of Fisher vectors [13] is to model the data with a generative model and compute the gradient of the likelihood of the data with respect to the parameters of the model, *i.e.* $\nabla_\lambda \log p(M|\lambda)$. We model M with a Gaussian mixture model (GMM) using Maximum Likelihood (ML) estimation. Let $\hat{u}_\lambda$ be the GMM model: $\hat{u}_\lambda(m) = \sum_{i=1}^{K} w_i u_i(\mu_i, \sigma_i)$, where K is the number of Gaussian components. The parameters of the models are $\lambda = \{w_i, \mu_i, \sigma_i, i = 1, \ldots, K\}$, where w_i denotes weight of the i-th component, while μ_i and σ_i are its mean and its standard deviations. We assume the covariance matrices are diagonal and σ_i represent the vector of standard deviations of the i-th component of the model. It worth pointing out that, considering the computational efficiency, for each image in the training set, only a randomly selected subset of local features is used to train the GMM model.

After getting the GMM, image representations are computed using Fisher vector, which is a powerful method for aggregating local descriptors and has been demonstrated to outperform the BoW model by a large margin [22].

Let $\gamma_t(i)$ be the soft assignment of the descriptor m_t to the component i:

$$\gamma_t(i) = \frac{w_i u_i(m_t)}{\sum_{j=1}^{K} w_j u_j(m_t)} \tag{2}$$

$G^M_{\mu,i}$ and $G^M_{\sigma,i}$ are the 7-dimensional gradients with respect to μ_i and σ_i of the component i. They can be computed using the following derivations:

$$G^M_{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i)\left(\frac{m_t - \mu_i}{\sigma_i}\right) \tag{3}$$

$$G^M_{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i)\left[\frac{(m_t - \mu_i)^2}{\sigma_i^2} - 1\right] \tag{4}$$

where the division between vectors is as a term-by-term operation. The final gradient vector G is the concatenation of the $G^M_{\mu,i}$ and $G^M_{\sigma,i}$ vectors for $i = 1, \ldots, K$ and is therefore $2 \times 7 \times K$-dimensional.

LDFV on color images. Previous works have shown that using color is a useful cue for person re-identification. We use the color information by splitting the image into 3 color channels (HSV), extract the proposed descriptor on each channel separately and finally concatenate the 3 descriptors into a single signature.

Similarity between LDFV representations Finally, the distance between two images I_i and I_j can obtained by computing the Euclidean distance between their representations :

$$d(I_i, I_j) = ||LDFV_i - LDFV_j|| \tag{5}$$

2.2 LDFV Extensions

bLDFV: using spatial Information.

To provide a rough approximation of the spatial information, we divide the image into many rectangular bins and compute one LDFV descriptor per bin. Please note that for doing this we compute one GMM per bin. Then, the descriptors of the different bins are concatenated to form the final representation. It is denoted by bLDFV, for bin-based LDFV.

It must be pointed out that our method does not use any body part segmentation. However, adapting the bins to body parts would be possible and could make the results even better.

eLDFV: combining LDFV with other features. As mentioned in the introduction, combining different types of image descriptors is generally useful. In this paper, we combine our bLDFV descriptor with two other descriptors: the Weighted Color Histograms (wHSV) and the MSCR, shown to be efficient for this task [1]. We denote this combination as eLDFV (enriched LDFV). In eLDFV, the difference between two image signatures $eD_1 = (HA_1, MSCR_1, bLDFV_1)$ and $eD_2 = (HA_2, MSCR_2, bLDFV_2)$ is computed as:

$$d_{eLDFV}(eD_1, eD_2) = \frac{1}{6} d_{wHSV}(HA_1, HA_2) + \frac{1}{6} d_{MSCR}(MSCR_1, MSCR_2) + \frac{2}{3} d_{bLDFV}(bLDFV_1, bLDFV_2); \tag{6}$$

Regarding the definition of d_{wHSV} and d_{MSCR}, we use the ones given in [1]. For simplicity reason and because it's not the central part of the paper, we have set the mixing weights by hand, giving more importance to the proposed descriptor. Learning them could certainly improve the results further.

sLDFV: using metric learning. In addition to the unsupervised similarity function (Eq. 5), we have also evaluated a supervised similarity function in which we use PCCA [20] to learn the metric. This variant is denoted sLDFV for supervised bLDFV. PCCA learns a projection into a low-dimensional space where the distance between pairs of data points respects the desired constraints, exhibiting good generalization properties in presence of high dimensional data. Please note that the bLDFV descriptors are pre-processed by applying a whitened PCA before PCCA. In sLDFV, PCCA is used with a linear kernel, for making the computations faster.

3 Experiment

The proposed approach has been experimentally validated on two person re-identification datasets (VIPeR [16] and ETHZ [3, 23]). After introducing the datasets, we present several experiments showing the efficiency of our simple LDFV descriptor and its extensions.

3.1 Datasets and Performance evaluation

The VIPeR [16] dataset contains $1,264$ images (normalized to 128×48 pixels) of 632 persons. There are exactly two images per person, taken from two non-overlapping viewpoints. As shown Fig. 1, VIPeR's images have a high degree of viewpoint and illumination variations: for most pairs there are 90 degrees from one of the other viewpoints. The VIPeR dataset has been widely used in the literatures, and is now considered as the benchmark of reference for person re-identification.

The ETHZ [3, 23] dataset contains three video sequences of crowded street scenes captured by two moving cameras mounted on a chariot. The three sequences are as follows: $4,857$ images of 83 pedestrians in SEQ. #1, $1,961$ images of 35 pedestrians in SEQ. #2, and $1,762$ images of 28 pedestrians in SEQ. #3. The most challenging aspects of ETHZ are illumination changes and occlusions.

For both datasets, the performance is measured by the Cumulative Matching Characteristic (CMC) curve [24], which is the standard metric for person re-identification. The CMC curve represents the probability of finding the correct match over the first n ranks.

3.2 Evaluation of the proposed image descriptor

In this section, our motivation is to evaluate the intrinsic properties of the descriptor. For this reason we don't use any metric learning but simply measure the similarity between two persons using the Euclidean distance between their representations.

Evaluation of the simple 7-d feature vector. The core of our descriptor is the 7-d simple feature vector given Eq. (1). This first set of experiments aims at validating this feature vector by comparing it with several alternatives, the rest of the framework being exactly the same. We did experiments with (i) SIFT features (reduced to 64 dimensions by PCA) and (ii) Gabor features [25] (with 8 scales and 8 orientations). For these experiments, we divide the bounding box into 12 bins (3×4) and the number of GMM components is set to 16. For each bin and each one of the 3 color channels (HSV), we compute the FV model and concatenate the 12 descriptors for obtaining the final representation. The size of the final descriptor is $7 \times 16 \times 12 \times 2 \times 3$ for our 7-d descriptor, $64 \times 16 \times 12 \times 2 \times 3$ for both the SIFT and Gabor descriptor based FV. We then compute CMC normalized Area under Curves (nAUC) on VIPeR get respectively 83.17, 86.37 and 91.60 for SIFT, Gabor and bLDFV using our 7-d feature vector. Consequently, the proposed descriptor, in addition of being compact and very simple to compute, gives much better results than SIFT and Gabor filters for this task.

We have evaluated the performance of our descriptor for different number of GMM components (16, 32, 50 and 100), and have observed that the performance is not very sensitive to this parameter. Consequently, we use 16 components in all of our experiments, which is a good tradeoff between performance and efficiency.

Fig. 1. VIPeR dataset: images of the same subjects from different viewpoint.

As set of representative images is required to learn the GMM, we conducted a set of experiments in order to evaluate how critical the choice for these images is. Out experiments have shown that using the whole dataset or only a smaller training set independent from the test set makes almost no difference, showing thereby that, in practice, a small set of representative images is more than enough for learning the GMM.

Single-shot experiments. Single-shot means that a single image is used as the query. We first present some experiments on the VIPeR dataset, showing the relative importance of the different components of our descriptor. As explained in Sec. 2 the full descriptor (eLDFV) is based on a basic Fisher encoding of the simple 7-d feature vector (LFDV) computed on the 3 color channels (HSV) and its two extensions, *i.e.* the spatial encoding (bLFDV) and the combination with two other features (namely wHSV and MSCR).

Fig. 2 gives the performance of eLDFV as well as the performance of wHSV, MSCR and bLDFV alone. We follow the same experimental protocol used in [1] and report the average performance over 10 random split of 316 persons. The figure also gives the performance of the state-of-the-art SDALF [1]. We can draw several conclusions: (i) LDFV alone performs much better than MSCR and wHSV (ii) using spatial information (bLFDV) improves the performance of LDFV (iii) combining the three components (eLDFV) gives a significant improvement over bLDFV and any of the individual components (iv) the proposed approach outperforms SDALF by a large margin. For example, the CMC score at rank 1, 10 and 50 for eLDFV are respectively of 22.34%, 60.04% and 88.82% while those of SDALF are of 19.84%, 49.37 and 84.84%.

We have also tested the proposed descriptor on the ETHZ database, in the single shot scenario ($N = 1$). Here again we follow the evaluation protocol proposed by [1]. Fig. 3 shows the CMC curves for the three different sequences. In the figure, dashed results come from [1]. The solid line are given by the proposed method. We can see that the performances of LDFV, bLDFV and eLDFV are all much better than that of the ones of SDALF, on all the three sequences, and

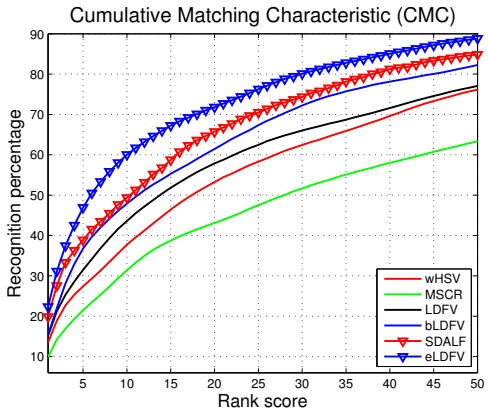

Fig. 2. VIPeR dataset: CMC curves of LDFV, bLDFV, eLDFV and SDALF.

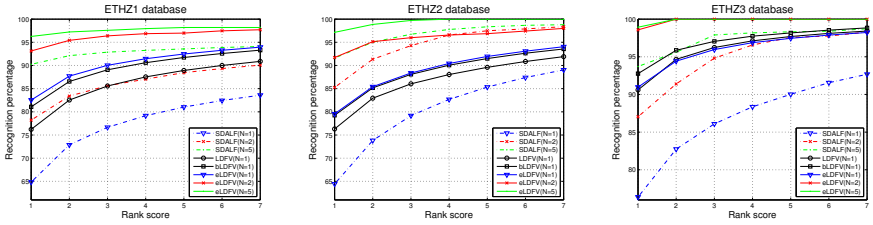

Fig. 3. CMC curves for the ETHZ dataset

improvements are even more visible than on VIPeR. Especially, on SEQ. 1 and 3, the performances of eLDFV are much worse than those of bLDFV though eLDFV is the combination of bLDFV, wHSV and MSCR. We attribute this to the low accuracy of wHSV and MSCR. In particular, on SEQ. 1, the minimum and maximum of the matching rate between the eLDFV and SDALF is about 10% and 18%, respectively. In SEQ. 2, rank 1 matching rate is around 80% for eLDFV and 64% for SDALF. The mean matching rate differences between eLDFV and SDALF on 7 ranks are about 10% in SEQ. 3.

Multi-shot experiments on ETHZ. Besides the single-shot case, we also test our descriptors in the multi-shot case. In this case $N \geq 2$ images are used as queries. We also follow the evaluation framework proposed by [1], the number of query images N being set to 2 and 5. Results are also shown Fig. 3. We can see that on SEQ. 1 and 3, eLDFV gives almost perfect results. Especially, on SEQ. 3, the performance of eLDFV is 100% with $N = 2, 5$, for ranks greater than 2.

3.3 Comparison with recent approaches

In this section we compare our framework with recent approaches. For making comparison fair, we use here the metric learning algorithm described Sec. 2.2.

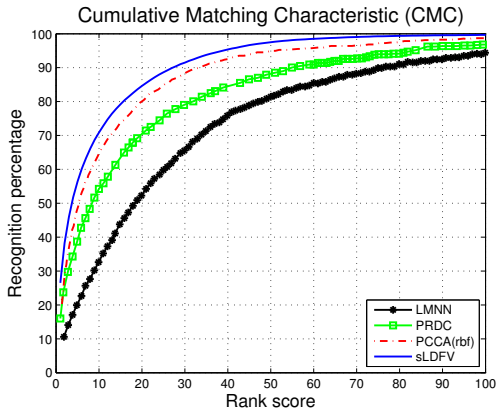

Fig. 4. VIPeR dataset: CMC curves with 316 persons.

We first present some experiments done on the VIPeR dataset. Following the standard protocol for this dataset, the dataset is split into a train and a test set by randomly selecting 316 persons out of the 632 for the test set, the remaining persons being in the train set. Like in [20], one negative pairs is produced for each person, by randomly selecting one image of another person. We produce 10 times more negative pairs than positive ones. The process is repeated 100 times and the results are reported as the mean/std values over the 100 runs.

Fig. 4 and Tab. 1 compare our approach (sLDFV) with three different approaches using metric learning: PRDC [26], LMNN [17] and PCCA [20]. The results of PRDC and LMNN are taken from [26] while the ones of PCCA come from [20]. For PRDC and LMNN, the image representation is the combination of RGB, YCbCr and HSV color features and two texture features extracted by local derivatives and Gabor filters on 6 horizontal strips. For PCCA, the feature descriptor is a 16 bins color histograms in 3 color spaces (RGB, HSV and YCrCb) as well as texture histograms based on Local Binary Patterns (LBP) computed on 6 non overlapping horizontal strips. PCCA [20] reports state-of-the-art results for person re-identification, improving over Maximally Collapsing Classes [27], ITML [18] or LMNN-R [28].

Fig. 4 and Tab. 1 shows that the proposed approach (sLDFV) performs much better than any previous approaches. For example, if we compare sLDFV with PCCA, we can see that matching rates for rank 1, 10 and 20 are of 26.53%, 70.88% and 84.63% for sLDFV while those of PCCA are only 19.27%, 64.91% and 80.28%. It must be pointed out that sLDFV is not using any non-linear kernel, from which we can expect further improvements.

4 Conclusion

In this paper, we have addressed the problem of person re-identification by proposing a novel descriptor, which is based on a simple seven-dimensional fea-

Table 1. VIPeR dataset: Top ranked matching rates (%) with 316 persons.

Method	r=1	r=5	r=10	r=20
PRDC [26]	15.66	38.42	53.86	70.09
MCC[26]	15.19	41.77	57.59	73.39
ITML[26]	11.61	31.39	45.76	63.86
LMNN[26]	6.23	19.65	32.63	52.25
CPS [11]	21.00	45.00	57.00	71.00
PRSVM [4]	13.00	37.00	51.00	68.00
ELF [6]	12.00	31.00	41.00	58.00
PCCA-sqrt n^{-}=10 [20]	17.28	42.41	56.68	74.53
PCCA-rbf n^{-}=10 [20]	19.27	48.89	64.91	80.28
sLDFV n^{-}=10	**26.53**	**56.38**	**70.88**	**84.63**

ture representation and the Fisher vector method. We test our descriptor on two challenging public datasets (VIPeR and ETHZ), outperforming the current state-of-the-art performance on both datasets.

There are several aspects to be further improved in the future, such as the weights of different features which are fixed by hand at the moment and should be learnt, or the seven-dimensional feature, based on pixels intensities, which can be made more robust to noise.

Acknowledgments

This work is a part of the Quaero Program funded by OSEO, French State agency for innovation and by the ANR, grant reference ANR-08-SECU-008-01/SCARFACE. The first author is partially supported by National Natural Science Foundation of China under contract No. 61003103.

References

1. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2010)
2. Oreifej, O., Mehran, R., Shah, M.: Human identity recognition in aerial images. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2010)
3. Schwartz, W., Davis, L.: Learning discriminative appearance based models using partial least squares. In: Brazilian Symp. on Comp. Graphics and Im. Proc. (2009)
4. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC. (2010)
5. Zhang, Y., Li, S.: Gabor-LBP based region covariance descriptor for person re-identification. In: Int. Conf. on Image and Graphics. (2011) 368 – 371
6. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proc. European Conference on Computer Vision. (2008)

7. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using haar-based and DCD-based signature. In: Proc. Int. Workshop on Activity Monitoring by Multi-camera Surveillance Systems. (2010)
8. Gheissari, N., Sebastian, T., Tu, P., Rittscher, J., Hartley, R.: Person re-identification using spatiotemporal appearance. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2006)
9. Kai, J., Bodensteiner, C., Arens, M.: Person re-identification in multi-camera networks. In: Proc. IEEE CVPR Workshops. (2011)
10. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: BMVC. (2009)
11. (Dong Seon Cheng, Marco Cristani, M.S.L.B., Murino, V.)
12. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2003)
13. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2007)
14. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. European Conference on Computer Vision. (2010)
15. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2010)
16. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. (2007)
17. Weinberger, K., Saul, L.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10** (2009) 207–244
18. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proc. International Conference on Machine Learning. (2007) 209–216
19. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: Proc. IEEE International Conference on Computer Vision. (2009)
20. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2012)
21. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Trans. on PAMI **30** (2008) 1713–1727
22. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC. (2011)
23. Ess, A., Leibe, B., Schindler, K., , van Gool, L.: A mobile vision system for robust multi-person tracking. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2008)
24. Moon, H., Phillips, P.: Computational and performance aspects of PCA-based face-recognition algorithms. Perception **30** (2001) 303–21
25. Fisher, R.A.: The use of multiple measures in taxonomic problems. Ann. Eugenics **7** (1936) 179–188
26. Zheng, W., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition. (2011)
27. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Advances in Neural Information Processing Systems. (2006)
28. Dikmen, M., Akbas, E., Huang, T., Ahuja, N.: Pedestrian recognition with a learned metric. ACCV (2010)