Matching People across Camera Views using Kernel Canonical Correlation Analysis

Giuseppe Lisanti , Iacopo Masi , Alberto Del Bimbo
{giuseppe.lisanti,iacopo.masi,alberto.delbimbo}@unifi.it
Media Integration and Communication Center (MICC), Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy

ABSTRACT

Matching people across views is still an open problem in computer vision and in video surveillance systems. In this paper we address the problem of person re-identification across disjoint cameras by proposing an efficient but robust kernel descriptor to encode the appearance of a person. The matching is then improved by applying a learning technique based on Kernel Canonical Correlation Analysis (KCCA) which finds a common subspace between the proposed descriptors extracted from disjoint cameras, projecting them into a new description space. This common description space is then used to identify a person from one camera to another with a standard nearest-neighbor voting method. We evaluate our approach on two publicly available datasets for re-identification (VIPeR and PRID), demonstrating that our method yields state-of-the-art performance with respect to recent techniques proposed for the re-identification task.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms, Experimentation, Measurement

Keywords

Person re-identification, KCCA

1. INTRODUCTION

Nowadays with the proliferation of cameras in airports and cities, a key component in modern and distributed surveillance systems is how to organize and search for soft biometrics of people. A soft biometric characteristic, that has recently emerged in computer vision, is the whole imaged body of the person. For this reason, a system that is able to search over a large database of people imagery[1] captured from different, non-overlapping distributed cameras, could be an helpful toolkit in the task of searching for an individual in a tangle of networked cameras of an airport. An algorithm to match people across camera views could be used in modern AOCC (Airport Operation Control Center). For example, it could be useful when an operator is asked to search for the identity in a gallery of hundreds of thousands of persons. This scenario is represented in Fig. 1, where the task is to assign a label to the probe image $\mathbf{p}_B$ considering all the gallery labels present in $\mathbf{G}_A$.

However, two non overlapping cameras usually contain a viewpoint change and a stark difference in illumination, background and camera characteristics that render the task of re-identification very challenging. The problem is even harder if we consider that we can have just one still image to describe an individual. This case is also known as Single versus Single modality (SvsS). Assuming that a training set is available, the aim of the paper is that of overcoming these issues in the single-shot modality in order to build a method that helps in tedious task of manually sifting through all the person imagery. Considering these issues, the main contributions of the paper are the following:

- we address the problem of person re-identification by

Figure 1: Person re-identification scenario: given an unlabeled probe image of a person $\mathbf{p}_B$ taken from camera B, find the corresponding person in the gallery set $\mathbf{G}_A$, containing people framed from camera A. The matched person is highlighted in green. Images are taken from the VIPeR dataset.

[1]In the rest of the paper, we will refer to person imagery as the whole imaged body silhouette, like the ones in Fig. 1

ICDSC '14 , November 04 - 07 2014, Venezia Mestre, Italy

http://dx.doi.org/10.1145/2659021.2659036

proposing an efficient but robust descriptor to encode the appearance of a person and by computing an exponential χ^2 kernel on top of this;

- the matching is improved by exploiting Kernel Canonical Correlation Analysis (KCCA) to find a common subspace between the proposed descriptors extracted from two disjoint cameras;
- to the best of our knowledge, we are the first to propose the use of Canonical Correlation Analysis to solve the ambiguity of representing people from different, disjoint cameras; in particular, usually CCA in literature is used to merge multi-modal data such as visual, text or tags [3, 18]; in this work, we show that KCCA could also be used to solve the camera transfer learning problem;
- despite the efficiency and compactness of our person representation, that is convenient for distributed cameras, our approach obtain state of the art performance compared to recent methods on two publicly available datasets widely used in literature, namely VIPeR [8] and PRID [12].

The paper is organized as follows: in Sec. 2 we review the most recent papers mainly focusing on the methods that address the re-identification as a metric learning problem; while in Sec. 3 we briefly give an overview of our method. In Sec. 3.1 we describe our descriptor and the kernels used; then in Sec. 3.3, we show how to use the KCCA in the re-identification problem. Finally in Sec. 4 we evaluate our method, based on KCCA, with respect to regular baselines such as standard nearest neighbor in the descriptor space (NN) and nearest neighbor in a linear space learned with CCA (CCA). Then we also compare our result with the most recent supervised techniques [22, 10, 11, 2, 12] and unsupervised ones [19, 21].

2. RELATED WORK

Recent works to solve the re-identification problem have moved from proposing hand-engineered features, that properly represent the appearance, to the task of camera transfer learning. Despite the effort of descriptor-based methods, lately, person re-identification has been casted as a metric learning problem usually parametrised as Mahalanobis distance. In fact, authors in [6] employ a metric learning algorithm to compute a robust Mahalanobis matrix using Large Margin Nearest Neighbor classification with Rejection (LMNN-R). The first authors to show that the person re-identification can be interpreted as a ranking problem are those in [17]: Prosser *et al.* reformulate the task as a ranking problem by learning, similar to us, a subspace where the potential true positive is given highest ranking rather than any direct distance measure such as ℓ_2 norm or similar. They develop an novel Ensemble RankSVM that maintains high-level performance and is able to overcome the scalability problems suffered by existing SVM-based ranking methods or SVM "one versus all" procedure. A recent generic metric learning approach is that proposed by [15] which also learns a Mahalanobis distance from equivalence constraints in a simple and straightforward manner that scales better to large dataset. The Probabilistic Distance Comparison (PRDC) approach [22] introduces a novel comparison model for people that maximizes the probability of a pair of correctly matched images to have a smaller distance than that of an incorrectly matched pair. They show that the proposed model is more tolerant to intra/inter-class variations that typically occur with multi-dimensional features.

Recently camera transfer approaches have been proposed to learn a metric parametrised differently than a Mahalanobis one [12] or using implicit learning [2]. The method in [12] encodes the person appearance extracting color (HSV and Lab) and texture information (LBP). This method, that shares some aspects of our approach, shows that when a linear metric has been learned properly, even a simple nearest neighbor classification technique can achieve compelling performance. A different framework proposed in [2], models camera transfer learning implicitly by learning a binary non-linear classifier with concatenations of pairs of vectors. The first pair describes an instance associated with camera A, and the second an instance associated with camera B. The classifier is learned in a supervised fashion constructing positive and negative labels considering the available training set. Similarly to [2], Martinel *et al.* [16] propose an approach that exploits pairwise dissimilarities between feature vectors to perform the re-identification in a supervised learning framework. The authors extract multiple features from two different cameras, compare them using standard distance metrics to obtain a distance feature vector (DFV) given a pair of descriptors. Then they learn the set of positive and negative distance feature vectors using random-forest tree and perform the re-identification by classifying the distance feature vector.

In contrast to supervised metric learning techniques, lately also unsupervised methods have been considered in literature. In [19] R. Zhao *et al.* employ saliency in the problem of matching people across views. First, they apply adjacency constrained patch matching to build dense correspondence between image pairs, being able to handle misalignment errors caused by large viewpoint and pose variations. Then they learn human salience in an unsupervised manner. The same method has been extended in [21] to better handle the misalignment problem by exploiting the fact that matching patches with inconsistent salience brings penalty. In this latter work, images of the same person are recognized by minimizing the salience matching cost. Conditional Random Fields have been exploited in [14] relying on a nearest neighbors topology between all the images of a dataset. The same authors also proposed a semi-supervised techniques where the data cost potential is estimated from SVM scores [13].

3. PROBLEM FORMULATION

Our approach consists of three steps that can be summarized in the following. Firstly we encode all the individuals using the descriptor proposed in Sec. 3.1. These descriptors are mapped through an exponential χ^2 kernel. Then we use the training data provided by each dataset to learn a common representation of the kernel descriptors from different cameras. Finally the projected kernel descriptors are evaluated in a common subspace using cosine distance. The whole process is shown in Fig. 2.

3.1 Person Representation

Considering an input image $I(x, y)$ containing a person, we resize the image to width and height respectively of $w = 64$ and $h = 128$ pixels. Then our approach slices the

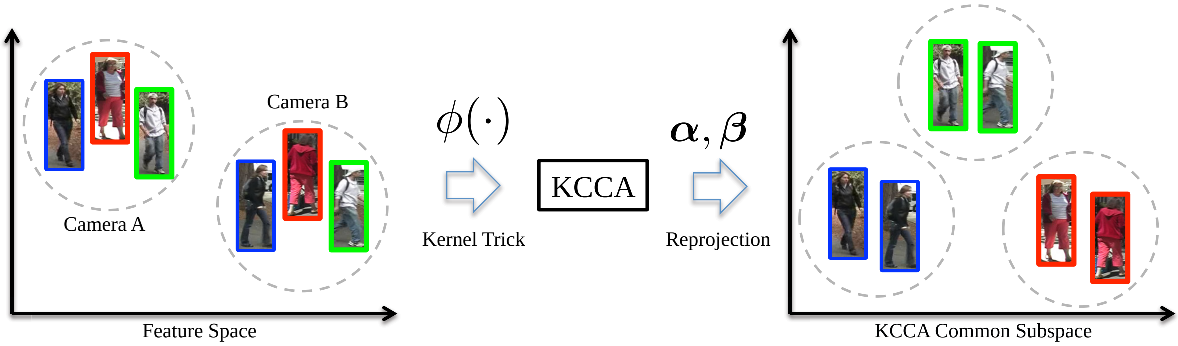


Figure 2: Overview of our method: feature descriptors in Euclidean space are mapped with the kernel trick in a higher dimensional space; then a common representation for the two cameras is learned on a training set through Kernel Canonical Correlation Analysis.

image in parts defined with horizontal stripes. From each stripe we extract three weighted color histograms in the Hue Saturation domain (discarding the Value information), in the RGB color space and in a color-opponent color space (Lab). The contribution of each pixel to each histogram bin is weighted through a non-isotropic Gaussian kernel centered in the image in order to decrease background pixel influence, such as:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \exp\Big(-\big(\frac{x-\mu_x}{\sigma_x^2} + \frac{y-\mu_y}{\sigma_y^2}\big)\Big). \tag{1}$$

We set the parameters of this kernel as $\boldsymbol{\mu} = [\mu_x, \mu_y] = [w/2, h/2]$, where w, h are respectively the width and height of the image. We found good experimental results by setting σ_x, σ_y as $w/4$ and $h/4$ pixels.

Each color histogram is concatenated to form the first part of the descriptor. At the end of this, we firstly add a HOG descriptor [4] quantizing the gradient orientations in 4 bin instead of 8 and, secondly, a texture analysis based on Local Binary Pattern (LBP) extracted both on a reduced region of the image centred on the torso and legs. LBP are extracted using the approach in [1], sampling LBP histogram on a grid with cell size egual to 16 pixels. Each LBP histogram is quantized in 58 bins, where two bins account for non-uniform binary patterns and the remaining count the term frequency of each uniform pattern.

This descriptor is extracted for all the available images of both camera A and camera B, such as:

$$\mathbf{D}_A = \underbrace{\left[\mathbf{d}_A^1\ \mathbf{d}_A^2\ \cdots\ \mathbf{d}_A^N\right]}_{\text{views from camera A}}, \mathbf{D}_B = \underbrace{\left[\mathbf{d}_B^1\ \mathbf{d}_B^2\ \cdots\ \mathbf{d}_B^N\right]}_{\text{views from camera B}}$$

where N is the number of subjects in the dataset. For each trial the two set $\mathbf{D}_A$ and $\mathbf{D}_B$ are randomly splitted into four subsets, two for each camera:

$$\mathbf{D}_A = [\mathbf{T}_A \mathbf{G}_A], \mathbf{K}_B = [\mathbf{T}_B \mathbf{P}_B] \tag{2}$$

where:

$$\mathbf{T}_A = \left[\mathbf{t}_A^1\ \mathbf{t}_A^2\ \cdots\ \mathbf{t}_A^{N_T}\right], \mathbf{T}_B = \left[\mathbf{t}_B^1\ \mathbf{t}_B^2\ \cdots\ \mathbf{t}_B^{N_T}\right]$$

represent the two training set from the two cameras, while:

$$\mathbf{G}_A = \left[\mathbf{g}_A^1\ \mathbf{g}_A^2\ \cdots\ \mathbf{g}_A^{N_G}\right], \mathbf{P}_B = \left[\mathbf{p}_B^1\ \mathbf{p}_B^2\ \cdots\ \mathbf{p}_B^{N_P}\right]$$

represent respectively the gallery $\mathbf{G}_A$ and the probe set $\mathbf{P}_B$.

3.2 Kernel Representation

We exploit the kernel trick to map our descriptor into a higher-dimensional feature space: $K(\mathbf{d}^i, \mathbf{d}^j) = \phi(\mathbf{d}^i)'\phi(\mathbf{d}^j)$. In particular, we use the χ^2 exponential kernel as:

$$K^{\chi^2}(\mathbf{d}^i, \mathbf{d}^j) = \exp\left(-\frac{1}{2C}\sum_k \frac{(\mathbf{d}^i - \mathbf{d}^j)^2}{(\mathbf{d}^i + \mathbf{d}^j)}\right), \tag{3}$$

where C is the median of the χ^2 distances among all the examples and the summation runs over the dimensionality of our feature descriptor $\mathbf{d}$. After applying the kernel trick we have,

$$\mathbf{K}_A = \begin{bmatrix} \mathbf{K}_A^{TT} & \mathbf{K}_A^{TG} \\ \mathbf{K}_A^{GT} & \mathbf{K}_A^{GG} \end{bmatrix}, \mathbf{K}_B = \begin{bmatrix} \mathbf{K}_B^{TT} & \mathbf{K}_B^{TP} \\ \mathbf{K}_B^{PT} & \mathbf{K}_B^{PP} \end{bmatrix}, \tag{4}$$

where the sub-matrices $\mathbf{K}_A^{TT}$ and $\mathbf{K}_B^{TT}$ represent the kernel version of our descriptor for the two training sets while the sub-matrices $\mathbf{K}_A^{GT}$ and $\mathbf{K}_A^{PT}$ represent the kernel version for the gallery and probe sets respectively, as defined in Eq. (2).

3.3 Matching People using KCCA

Given the two views of the data projected as described in Sec. 3.2 we can construct a common representation exploiting the labeled training data.

3.3.1 *Canonical Correlation Analysis*

The Canonical Correlation Analysis (CCA) constructs the subspace that maximizes the correlation between two paired variables. More formally, given N_T training samples from a paired dataset, that in our case are the views of the data from two different non-overlapping cameras:

$$\mathbf{T} = \left\{(\mathbf{t}_A^1, \mathbf{t}_B^1), (\mathbf{t}_A^2, \mathbf{t}_B^2), ..., (\mathbf{t}_A^{N_T}, \mathbf{t}_B^{N_T})\right\} \tag{5}$$

the aim is to solve:

$$\rho = \max_{\mathbf{w}_A, \mathbf{w}_B} \operatorname{corr}(\mathbf{w}_A \mathbf{T}_A, \mathbf{w}_B \mathbf{T}_B) \tag{6}$$

in order to maximize the correlation between the two projected sets of points, $\mathbf{w}_A \mathbf{T}_A$ and $\mathbf{w}_B \mathbf{T}_B$.

3.3.2 *Kernel Canonical Correlation Analysis*

The Kernel Canonical Correlation Analysis (KCCA) performs as CCA but on data projected through an opportune kernel. As defined in Sec. 3.2, the kernel computed over $\mathbf{T}_A$

and $\mathbf{T}_B$ can be also expressed as:

$$K^{\chi^2}(\mathbf{t}_A^i, \mathbf{t}_A^j) = \phi(\mathbf{t}_A^i)'\phi(\mathbf{t}_A^j), \quad (7)$$

$$K^{\chi^2}(\mathbf{t}_B^i, \mathbf{t}_B^j) = \phi(\mathbf{t}_B^i)'\phi(\mathbf{t}_B^j). \quad (8)$$

Since $\mathbf{w}_A$ and $\mathbf{w}_B$ lie in the span of the N_T training instances, such as $\mathbf{w}_A \in span(\phi(\mathbf{T}_A))$ and $\mathbf{w}_B \in span(\phi(\mathbf{T}_B))$, we can re-write it for the KCCA as:

$$\mathbf{w}_A = \sum_i \boldsymbol{\alpha}_i \phi_A(\mathbf{t}_A^i),$$

$$\mathbf{w}_B = \sum_i \boldsymbol{\beta}_i \phi_B(\mathbf{t}_B^i),$$

where $i \in [1, \ldots, N_T]$. The objective of KCCA is thus to identify the weights $\boldsymbol{\alpha}, \boldsymbol{\beta}$ that maximize:

$$\arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{\boldsymbol{\alpha}'\mathbf{K}_A^{TT}\mathbf{K}_B^{TT}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}'\mathbf{K}_A^{TT}\mathbf{K}_A^{TT}\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{K}_B^{TT}\mathbf{K}_B^{TT}\boldsymbol{\beta}}}, \quad (9)$$

where $\mathbf{K}_A^{TT}$ and $\mathbf{K}_B^{TT}$ denote the $N_T \times N_T$ kernel matrices of the N_T sample pairs from the training set. As shown by Hardoon [9], learning may need to be regularized in order to avoid trivial solutions. Hence, we penalize the norms of the projection vectors and obtain the standard eigenvalue problem:

$$(\mathbf{K}_A^{TT} + \kappa\mathbf{Id})^{-1}\mathbf{K}_B^{TT}(\mathbf{K}_B^{TT} + \kappa\mathbf{Id})^{-1}\mathbf{K}_A^{TT}\boldsymbol{\alpha} = \lambda^2\boldsymbol{\alpha}. \quad (10)$$

The top M eigenvectors of this problem yield:

$$\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}^{(1)} \ldots \boldsymbol{\alpha}^{(M)}\right], \boldsymbol{\beta} = \left[\boldsymbol{\beta}^{(1)} \ldots \boldsymbol{\beta}^{(M)}\right]$$

that represent the semantic projections that will be used for both gallery and probe data.

3.3.3 Re-identification in the common subspace

At test time, we project the probe samples with $\boldsymbol{\alpha}$ and the gallery samples with $\boldsymbol{\beta}$:

$$\mathbf{G}_{\boldsymbol{\alpha}} = \mathbf{K}^{GT}\boldsymbol{\alpha}, \quad (11)$$

$$\mathbf{P}_{\boldsymbol{\beta}} = \mathbf{K}^{PT}\boldsymbol{\beta} \quad (12)$$

Then we compute the cosine distance between the projected descriptors of the gallery and probe and we perform a simple Nearest Neighbor (NN) classification, such that:

$$\text{id}(\mathbf{p}_{\boldsymbol{\beta}}) = \arg\min_i \left(\frac{\mathbf{g}_{\boldsymbol{\alpha}}^i \mathbf{p}_{\boldsymbol{\beta}}}{||\mathbf{g}_{\boldsymbol{\alpha}}^i||_2 ||\mathbf{p}_{\boldsymbol{\beta}}||_2}\right) \quad (13)$$

where i represent the identity of the i-th gallery sample.

4. EXPERIMENTAL RESULTS

In this section we evaluate our method with respect to two regular baselines such as standard nearest neighbor in the descriptor space using ℓ_2 norm (NN) and nearest neighbor in a subspace learned with CCA using cosine distance (Linear CCA). Then we also compare our result with the most recent supervised techniques such as PRDC [22], DDC [10], EIML [11], ICT [2], RPLM [12] and unsupervised ones like SalMatch [21] and eSDC [19]. We carry out our experiments on two widely used dataset for re-identification that are VIPeR [8] and PRID [12] considering CMC (Cumulative Matching Characteristic) curves.

We use the code provided by Hardoon *et al.* [9] as CCA and KCCA implementations. In our experiments we set the

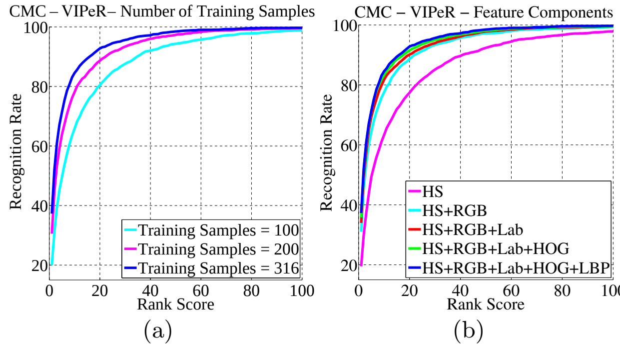


Figure 3: (a) CMC curves varying the number of pairs N_T in the training set. (b) CMC curves varying features components.

reconstruction error of the Partial Gram-Schmidt Orthogonalization (PGSO) as $\eta = 1$ while we set the regularisation parameter used in Eq. (10) as $\kappa = 0.5$ for both VIPeR and PRID datasets. These settings were carried out by dividing the training set in two parts, one for training and one for validation to estimate the best settings. Then we re-trained the KCCA on the whole training data using the best values of η and κ. The source code of our method is publicly available online[2].

4.1 VIPeR dataset

VIPeR dataset is the most challenging dataset currently available for person re-identification. It is composed by 632 image pairs of people captured outdoor (thus 1264 images in total), from two different, non-overlapping camera views. The challenges in VIPeR are mainly due to viewpoint and illumination variations, which cause severe appearance changes. We use the experimental protocol widely used in literature for metric learning techniques: the set of 632 image pairs is randomly split into two sets of 316 image pairs, one for training and one for testing. A single image from the probe set is then selected and matched with all the images from the gallery set. This process is repeated for all images in the probe set independently. The whole evaluation procedure is carried out on the 10 splits publicly available from [7]. Table 1 gives an overview of the recognition rate at various ranks of our approach compared with recent methods. We report an increment of 7% at first rank with respect to the SalMatch approach [21] that currently holds state-of-the-art performance. Considering supervised method, the best result is obtained by RPML [12]. However, we outperform RPLM and the others supervised approaches [22, 10, 11, 2] at all ranks on the VIPeR dataset. This performance arises from the combination of an efficient and robust descriptor that is additionally exploited by lifting the feature in a higher dimensional space using kernel trick and solving the camera transfer ambiguity via KCCA. Our approach saturates earlier than recent methods by reaching 93% of recognition rate at rank 20 while the other techniques report results in the range of 70 − 80%. Fig. 4(a) shines more light on this and demonstrates the steep trend of our CMC curve w.r.t the baselines.

Our approach with KCCA using the kernel of Eq. (3)

[2] http://www.micc.unifi.it/lisanti/source-code/

	VIPeR					PRID				
Rank:	*1*	*10*	*20*	*50*	*100*	*1*	*10*	*20*	*50*	*100*
LMNN [20]	17	54	69	88	96	–	–	–	–	–
ITML [5]	13	53	71	90	97	–	–	–	–	–
PRDC [22]	16	54	70	87	97	–	–	–	–	–
DDC [10]	–	–	–	–	–	4	24	37	56	70
EIML [11]	22	63	78	93	98	15	38	50	67	80
ICT [2]	14	60	78	–	–	–	–	–	–	–
RPLM [12]	27	69	83	95	99	**15**	42	54	70	80
eSDC [19]	27	62	76	–	–	–	–	–	–	–
SalMatch [21]	30	65	–	–	–	–	–	–	–	–
Proposed	**37**	**85**	**93**	**98**	**100**	**15**	**47**	**60**	**75**	**87**

Table 1: Comparative performance analysis at rank-1 with respect to baseline (first two rows) and the state-of-the-art on VIPeR and PRID datasets. Recognition rates in percentage.

	N_T=100			N_T=200			N_T=316		
Rank:	*1*	*10*	*20*	*1*	*10*	*20*	*1*	*10*	*20*
PRDC [22]	11	38	52	20	56	71	16	54	79
RPLM [12]	9	34	49	13	47	63	27	69	83
Proposed	**20**	**66**	**80**	**31**	**78**	**89**	**37**	**85**	**93**

Table 2: Recognition Rate in function of the number of pairs N_T in the training set.

outperforms both Nearest Neighbor classification in the features space (NN) and the linear version of Canonical Correlation Analysis (Linear CCA). It is possible also to see that employing the base descriptor with NN provides decent performance but is not achieving state-of-the-art results, especially at high ranks. Linear CCA slightly improves the recognition rate with respect to NN. On the VIPeR dataset we evaluate also how our method is sensitive to the number of training samples N_T, and we compare it with RPLM [12] and PRDC [22]. Table 2 shows the recognition rate across ranks with a different number of training samples $N_T \in \{100, 200, 316\}$. From these experiments, it is worth to notice that our approach is less sensitive to the number of training images with respect to [12] and [22] (see Fig. 3(a)). We also show in Fig. 3(b) the contribution of each component of the person representation proposed in Sect. 3.1. Here it is possible to appreciate that the performance is almost dominated by HS and RGB components while the others slightly improve the first rank recognition rate.

Finally, to show the effectiveness of the proposed approach with respect to metric learning ones, in Table 3 we give a comparison with two state of the art techniques [22, 15], by employing their descriptors in our method. We used these two metric learning methods since their descriptors are publicly available.

4.2 PRID dataset

The Person Re-ID (PRID) dataset consists of images extracted from multiple person trajectories recorded from two different, static surveillance cameras. This dataset was firstly introduced by [10] and is different from the other available datasets in that the number of gallery identities is higher than the probes ones. This means that the dataset has some distractors in the gallery set that are not associated

	VIPeR				
Rank:	*1*	*10*	*20*	*50*	*100*
PRDC [22]	16	54	70	87	**97**
Proposed (PRDC descriptor)	**20**	**64**	**78**	**91**	96
KISSME [15]	20	62	**81**	92	–
Proposed (KISSME descriptor)	**22**	**67**	80	**93**	**98**

Table 3: Performance comparison with metric learning approaches obtained by employing their descriptors in our method.

with probe individuals. In particular, while the view A has framed 385 persons, the view B has 749 persons. Considering these identities only the first 200 persons appear in both cameras. In our experiments we used the single-shot version of the dataset as in [12]. Considering this scenario, 100 identities are randomly chosen from the 200 present in both camera views for the training set while the remaining 100 of the first camera are used as probe set, and the 649 of the second view are used as gallery. The whole evaluation procedure is carried out on 10 splits. Since the dataset is more recent than VIPeR, we compare our method with the following approaches [10, 11, 12]. Table 1 summarizes the trend of the CMC curves at some selected ranks. Our approach obtains the same performance of RPLM [12] at rank-1 while at higher ranks starts to outperform all the other techniques. PRID dataset is more challenging than VIPeR for the presence of distractors and for the different viewpoint of the cameras. In fact, it is possible to observe in Fig. 4(b) that the performance largely improves while considering only 100 individuals in the gallery (without distractors) instead of considering all the 649 identities.

5. CONCLUSION

In this paper we have proposed a method to match people across views by learning a common subspace that reduces the ambiguity when descriptor are extracted from different disjoint cameras. Our method exploits Kernel Canonical Correlation Analysis (KCCA) to solve the camera transfer learning problem. Our approach demonstrates compelling performance in the re-identification task on two reference datasets used in literature.

Acknowledgments

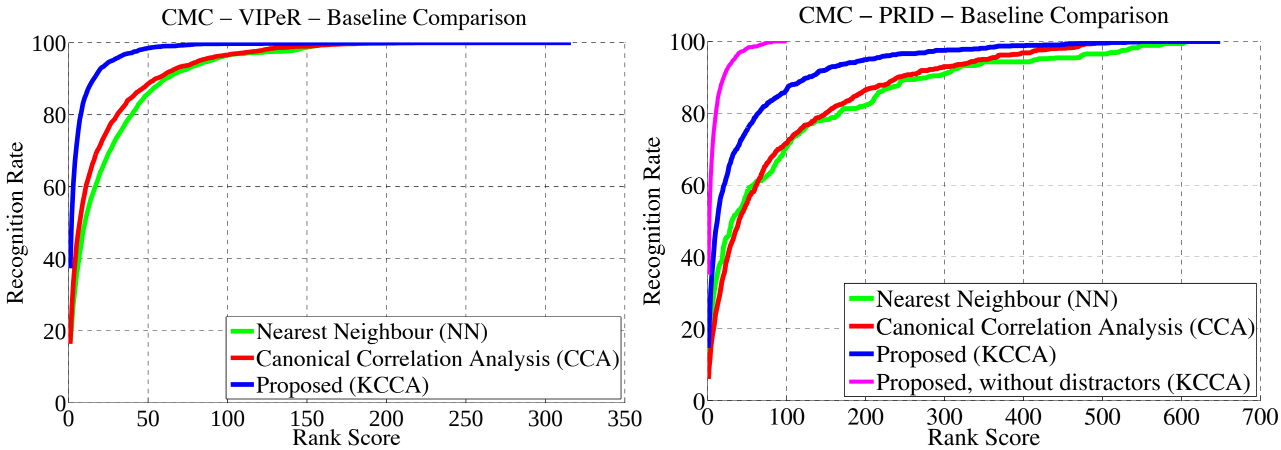
This work is partially supported by Thales Italia. We want to thank Tiberio Uricchio for his fruitful discussion and advices.

6. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *International Workshop on Re-Identification In conjunction with ECCV*, 2012.

[3] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *Proc. of International Conference on Multimedia Retrieval*, 2014.



(a) VIPeR (b) PRID

Figure 4: Baseline Comparison on VIPeR and PRID dataset: figures show the CMC of nearest neighbor in the feature space (green), in the space learned by CCA (red) and by the proposed KCCA (blue).

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2005.

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, Corvalis, Oregon, USA, June 2007.

[6] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Proc. of the Asian Conference on Computer Vision*, 2011.

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

[9] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conf. on Image Analysis*, 2011.

[11] M. Hirzer, P. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *IEEE 9th Int'l Conference on Advanced Video and Signal-Based Surveillance*, 2012.

[12] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012.

[13] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. D. Bimbo. Leveraging local neighborhood topology for large scale person re-identification. *Pattern Recognition*, 2014. In press.

[14] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. Del Bimbo. From re-identification to identity inference: Labeling consistency by local similarity constraints. In *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 287–307. 2014.

[15] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[16] N. Martinel, C. Micheloni, and C. Piciarelli. Learning pairwise feature dissimilarities for person re-identification. In *International Conference on Distributed Smart Cameras*, 2013.

[17] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010.

[18] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*, 2010.

[19] X. W. Rui Zhao, Wanli Ouyang. Unsupervised salience learning for person re-identification. In *Computer Vision and Pattern Recognition*, 2013.

[20] K. Weinberger and L. Saul. Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th international conference on Machine learning*, pages 1160–1167. ACM, 2008.

[21] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013.

[22] W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1, 2012.