Multi Channel-Kernel Canonical Correlation Analysis for Cross-View Person Re-Identification

Giuseppe Lisanti, University of Florence
Svebor Karaman, Columbia University
Iacopo Masi, University of Southern California

In this paper, we introduce a method to overcome one of the main challenges of person re-identification in multi-camera networks, namely cross-view appearance changes. The proposed solution addresses the extreme variability of person appearance in different camera views by exploiting multiple feature representations. For each feature, Kernel Canonical Correlation Analysis (KCCA) with different kernels is employed to learn several projection spaces in which the appearance correlation between samples of the same person observed from different cameras is maximized. An iterative logistic regression is finally used to select and weight the contributions of each projection and perform the matching between the two views. Experimental evaluation shows that the proposed solution obtains comparable performance on the VIPeR and PRID 450s datasets and improves on the PRID and CUHK01 datasets with respect to the state-of-the-art.

CCS Concepts: •**Computing methodologies** → **Computer vision tasks;**

Additional Key Words and Phrases: Person re-identification, KCCA, Late fusion

ACM Reference Format:
Giuseppe Lisanti, Svebor Karaman and Iacopo Masi. 2016. Multi Channel-Kernel Canonical Correlation Analysis for Cross-View Person Re-Identification. *ACM Trans. Multimedia Comput. Commun. Appl.* 0, 0, Article 0 (0), 19 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Video surveillance systems are now ubiquitous in public areas such as airports, train stations or even city wide. These systems are typically implemented in the form of camera networks and cover very large areas, with limited or no overlap between different camera views. They should be able to track a person throughout the network, matching detections of the same person in different camera views, irrespectively of view and illumination changes, as well as pose and scale variations of the person. Matching person detections across a camera network is typically referred as *re-identification*.

In this paper, we propose a solution for person re-identification that grounds on the idea of addressing the extreme variability of person appearance in different camera views through a multiplicity of representations. In particular, several color and texture features are extracted from a coarse segmentation of the person image to account for viewpoint and illumination changes. For each feature, we learn several projection

This research is partially supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST.
Authors' addresses: G. Lisanti is with the Media Integration and Communication Center (MICC) of the University of Florence, Florence, Italy. E-mail: giuseppe.lisanti@unifi.it; S. Karaman is with the Digital Video and Multimedia Lab, Department of Electrical Engineering of Columbia University, New York, NY, USA. E-mail: svebor.karaman@columbia.edu I. Masi is with the Institute for Robotics and Intelligent Systems of University of Southern California, Los Angeles, CA, USA. E-mail: iacopo.masi@usc.edu

DOI: 0000001.0000001

spaces where features computed on images of the same person observed in two different cameras correlate. These projection spaces are learned using Kernel Canonical Correlation Analysis (KCCA) with different kernels. Finally, matching between images from two cameras is performed by applying an iterative logistic regression procedure that enables selecting and weighting the contributions of the distances computed in each projection space.

1.1. Related works

Re-identification has been an active subject of research for several years, as recently surveyed in [Bedagkar-Gala and Shah 2014]. We review the most important works in the following. The approaches proposed in literature can be categorized in four categories: defining hand-crafted person descriptors, deep learning for person re-identification, learning discriminative models for person re-identification and learning a common space for person re-identification.

1.1.1. Hand-crafted person descriptors. These methods concentrate on the definition of descriptors that are able to capture as much as possible the variability of person appearance in different views. Approaches in this category often rely on the definition of regions of the image that should correspond to the different body parts of a person. Each region is usually encoded with color histograms or by the aggregation of local feature descriptors. Among the best performing proposals in this class, the Symmetry-Driven Accumulation of Local Features (SDALF) descriptor [Farenzena et al. 2010] takes into account image segments of physical parts of the human body such as the head, torso, and legs, obtained from the computation of axis symmetry and asymmetry and background modeling. For each segment, color information is represented by weighted HSV color histograms and Maximally Stable Color Regions (MSCR), and texture information is encoded as recurrent highly-structured patches. In [Cheng et al. 2011] the same authors proposed to fit a Custom Pictorial Structure (CPS) model on a person detection estimating the head, chest, thighs and legs positions. Each part is then described by a HSV color histogram and MSCR.

Many descriptor-based methods proposed in the literature are discussed in the survey [Doretto et al. 2011]. Furthermore, the authors of [Vezzani et al. 2013] review a large body of the research on re-identification with a focus on 2D and 3D model based approaches. However, most of the descriptor-based approaches in the literature rely on part-based models, and while performing well for ideal capture conditions, they have poor performance in real scenarios. This is due to the fact that image quality is often low and it is hard to precisely detect body parts.

1.1.2. Deep learning for person re-identification. As opposed to the design of hand-crafted features, some authors have exploited Deep Convolutional Neural Networks (CNN) to build a representation that captures the variability of person appearance across views. One of the first re-identification works in this class was [Yi et al. 2014a]. Successively, in [Yi et al. 2014b], the same authors improved their solution by employing a CNN in a "siamese" configuration to jointly learn the color feature, texture feature and the distance function in a unified framework (Improved DML). Ahmed *et al.* [Ahmed et al. 2015] proposed a siamese deep network architecture, similar to [Yi et al. 2014b], that learns jointly the feature representation and to discriminate between pairs of target in a same/not-same fashion (Siamese CNN) with a logistic regression loss. Finally, Li *et al.* [Li et al. 2014] used a novel filter pairing neural network (FPNN) with six-layers to jointly handle photometric and geometric transforms.

While deep learning has had a big impact on general image recognition and recently on face recognition [Taigman et al. 2014], the use of Deep Network-based representations for re-identification is negatively affected by low resolution images that usually

occur in re-identification contexts and requires the availability of a huge number of person image pairs from different cameras to train a discriminative model.

1.1.3. Learning discriminative models for person re-identification. This class of methods is the most populated and grounds on the idea of learning classifiers and metrics to recognize persons across views. They currently score the state-of-the-art performance of re-identification. In [Köstinger et al. 2012] the authors proposed a Mahalanobis based distance learning that exploits equivalence constraints derived from target labels (KISSME). The authors of [Hirzer et al. 2012a] proposed an impostor-based metric learning method (EIML), based on a modified version of the Large Margin Nearest Neighbor (LMNN) [Weinberger and Saul 2009] algorithm. The method in [Xiong et al. 2014] combined Regularized Pairwise Constrained Component Analysis, Kernel Local Fisher Discriminant Analysis, Marginal Fisher Analysis and a Ranking Ensemble Voting Scheme with linear, χ^2 and RBF-χ^2 kernels to extensively evaluate person re-identification performances (KLMM). Similarly to [Xiong et al. 2014], the approach in [Wang et al. 2016] introduced an explicit non-linear transformation for the original feature space and learned a linear similarity projection matrix (SLTRL) by maximizing the top-heavy ranking loss instead of a loss defined by the Area Under the Curve. Remarkable performance has been also obtained by [Paisitkriangkrai et al. 2015; Liu et al. 2015b]. The former combined an ensemble of different distance metric learning approaches, minimizing different objective functions, while the latter proposed a novel ensemble model (ECM) that combines different color descriptors through metric learning. The authors of [Liu et al. 2015a] proposed the Kernelized Relaxed Margin Components Analysis (KRMCA) approach that learns a metric exploiting both the nearest true neighbors and impostors during training.

The methods proposed in [Rui Zhao 2013; Zhao et al. 2013; Zhao et al. 2014] rely mainly on dense correspondences and unsupervised learning of features. In [Rui Zhao 2013], a novel method (eSDC) was proposed that applies adjacency-constrained patch matching to build dense correspondences between image pairs through a saliency learning method in a unsupervised fashion. The authors of [Zhao et al. 2013] extended this method by penalizing patches with inconsistent saliency in order to handle misalignment problems (SalMatch). Finally, instead of relying on hand-crafted features, Zhao *et al.* [Zhao et al. 2014] proposed to learn mid-level filters (mFilters). Dense patches are clustered together in order to create a hierarchical tree, then the patches inside a node of the tree are used to train a linear SVM that discriminates patches of the two views. Here, the mFilters are represented by the set of SVM weights and biases learned over the nodes. Differently from [Zhao et al. 2014], the method in [Shen et al. 2015] introduced a structure to encode cross-view pattern correspondences (CSL) that are used jointly with global constraints to exclude spatial misalignments.

The method in [de Carvalho Prates and Schwartz 2015] used salient samples from the probe and the gallery to build a set of prototypes. These prototypes are used to weight the features according to their discriminative power by using Partial Least Square (PLS). The final recognition is performed by fusing different rank results. In [Yang et al. 2014] the authors proposed to encode color using color naming. In particular, color distributions over color names in different color spaces are fused to generate the final feature representation (SCNCD). This method employs the KISSME metric learning framework to perform matching. The work [Shi et al. 2015] proposed to address the person re-identification problem by leveraging semantic attributes. The main underlying idea is that attributes may provide a strong invariant cue for recognition. Instead of relying on manually labeled attributes, the model is trained on fashion photography data. The attributes are learned as latent variables on top of a superpixel rep-

resentation. The authors also transferred the learned model to the video-surveillance setting without requiring any surveillance domain supervision.

1.1.4. Learning a common space for person re-identification. Even though the key problem in re-identification is to mitigate the strong appearance variations a subject undergoes across cameras, there are only few methods that tackle directly this issue. In particular, techniques that deal with cross-view matching by learning a common feature space to remove appearance changes across views were proposed in [An et al. 2013; Lisanti et al. 2014; An et al. 2015; An et al. 2016; Li and Wang 2013; Liao et al. 2015]. These methods are the closest to our approach since they learn feature projections to better perform matching between images of the same person captured from different cameras.

In [An et al. 2013], the authors were the first to apply CCA (Canonical Correlation Analysis) successfully to the re-identification problem. In particular, they apply regularized CCA offline between the gallery set and a reference set. Probes are then projected into the same common space and matched using cosine similarity. Successively, in [Lisanti et al. 2014], the authors extended CCA into its kernelized version and obtained remarkable results. In [An et al. 2015] linear CCA with a robust estimation of the covariance matrix is used to deal with small training sets (ROCCA). This solution leads to better accuracy compared to the regular CCA. The same authors extended their method in [An et al. 2016] to incorporate a reference set for each camera view, instead of applying the CCA to the gallery set only. The reference set is a group of disjoint persons from the test sets, that is used as outside data to compare probe and gallery. In [Li and Wang 2013] the authors proposed to use the similarity of cross-view transformations to partition the image spaces of two camera views into different configurations. Then, the visual features of an image pair from different views are projected into a common feature space and matched with softly assigned metrics. A discriminative metric is also learned to better discriminate between subjects. The authors of [Liao et al. 2015] defined the LOMO feature which is composed by HSV color histograms over stripes along with a texture descriptor which improves over the classic LBP. Their approach revised the KISSME metric learning [Köstinger et al. 2012] in order to deal with cross-view matching problem as well. The final proposed metric is named XQDA.

1.2. Contributions and Distinctive Features

Our approach grounds on many person descriptors proposed in the literature [Prosser et al. 2010; Karaman and Bagdanov 2012; Lisanti et al. 2015; Karaman et al. 2014; Liao et al. 2015] which used a coarse spatial segmentation of the image into consecutive regions at different heights. Each region is represented by multiple features to capture the variety of a person appearance. Our work largely differs from the widespread effort in the community employing linear [Li and Wang 2013; Liao et al. 2015] and also non-linear metric-learning [Liu et al. 2015a]. Indeed, the core of our approach is to learn multiple representations through non-linear CCA for each feature, while other works used its linear version [An et al. 2013; An et al. 2015] on a single descriptor.

Considering this, the two major novel contributions of the proposed work are the following:

— For each feature, we learn a set of projection spaces with different kernels, such that images of the same person coming from different cameras are more easily matched. This differs from the approaches of [Li and Wang 2013; Lisanti et al. 2014; Liao et al. 2015] where a single projection space was learned and from [An et al. 2013; An et al. 2015] that are based on linear CCA.

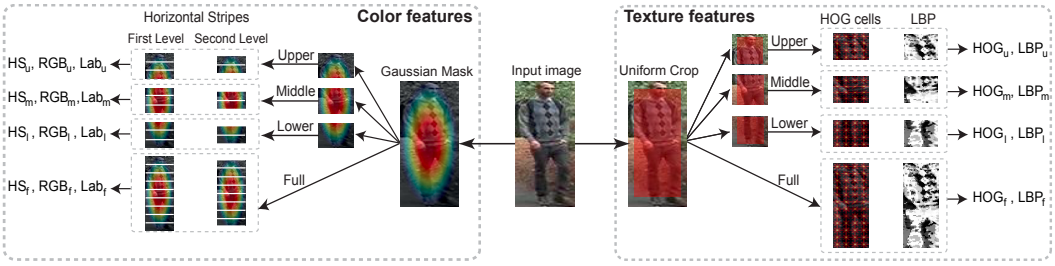


Fig. 1. Illustration of our feature descriptor extraction process. We extract color (HS, RGB and Lab) and texture (HOG and LBP) features from the full image and from the upper, middle and lower regions of the image.

— We derive an iterative selection procedure, based on logistic regression, where less significant features are dropped out and distinguishing features contribute more to the re-identification. This allows us to improve the re-identification performance while reducing the computational cost at test time.

In addition to those contributions, we also publicly release our code[1] to promote the reproducibility of our results and enable the community to further build upon our work.

In the rest of the paper, we expound our person representation in Sect. 2, and discuss in detail the method in Sect. 3 and 4. In Sect. 5, we compare performance for re-identification using KCCA with multiple kernels with respect to methods learning a common space or using metric learning. We also give an overview of the performance of our method with respect to the state-of-the-art on person re-identification. Finally, we show the contribution of each feature and kernel used in our solution, we give some insights on our iterative selection procedure and discuss the computational cost of our method.

2. PERSON REPRESENTATION

In order to account for spatial distribution of the person appearance, our representation model considers four components: the full person image and a coarse segmentation into upper, middle and lower regions. Color information is extracted from each component and modeled by histograms in the Hue Saturation, RGB and Lab color spaces, in order to account for differences in illumination due to different viewpoints. Texture features are also extracted for each component and represented with HOG [Dalal and Triggs 2005] and Local Binary Pattern (LBP) histograms. Therefore, for each component we extract multiple features, namely: HS_p, RGB_p, Lab_p, HOG_p, LBP_p, where the suffix p stands for the full (f), upper (u), middle (m) and lower (l) components of our representation. Later on, we dub a feature extracted in one component as a channel and denote $\mathcal{C}$ the set of channels.

The features extraction process goes as follows, the person images are first resized to the resolution of 126×64 pixels. For color features, the contribution of each pixel to each histogram bin is weighted through a non-isotropic Gaussian kernel to decrease background pixels influence without requiring an explicit background segmentation. Furthermore, a segmentation into overlapping stripes of 14 pixels is considered for each component (see Fig. 1). For one component, each color histogram is computed for each stripe using 64 bins and concatenated across stripes. The HS_f, RGB_f and Lab_f have thus a dimensionality of 1088 (17×64), while color features for each upper, mid-

[1] https://github.com/glisanti/MCK-CCA

dle and lower regions have 320 dimensions (5×64). Regarding these color descriptors, the parameters used to process a person image are mostly derived from [Lisanti et al. 2015], except for some minor variations in the stripes configuration and the additional Lab color space we introduce in this work. For texture channels, we remove 6 pixels from the image border and compute the HOG descriptor using 4 bins for the gradient orientations. The HOG_f has 1040 dimensions while each region feature has 320 dimensions. Differently from [Lisanti et al. 2015], we also added another texture descriptor based on LBP using the standard quantization of LBP histograms proposed in [Ojala et al. 2002]. More precisely, the LBP features are computed on a grid with cells of 16 pixels using 58 bins. The LBP_f has 1218 dimensions while each region feature has 348 dimensions.

3. MULTI CHANNEL-KERNEL CANONICAL CORRELATION ANALYSIS

Matching two images of the same person coming from two different cameras can be difficult due to illumination changes and pose variations. As different features and component of the image can be affected differently, we propose to learn a common projection space for each channel (one feature extracted in one component) separately. To learn these common projection spaces, we employ KCCA [Hardoon et al. 2004] which has been shown to be effective when applied to a whole image descriptor, as in [Lisanti et al. 2014].

We introduce the following notation. Given a feature channel $c \in \mathcal{C}$, let us denote $\mathbf{F}_a(c)$ the set of feature vectors $\mathbf{f}_a(c)$ and $\mathbf{F}_b(c)$ the set of feature vectors $\mathbf{f}_b(c)$, respectively for camera a and b, and using camera a for gallery and camera b for probe, we define:

$$\mathbf{F}_a(c) = \left[\mathbf{F}_a^T(c) \mid \mathbf{F}_a^G(c) \right] \tag{1}$$

$$\mathbf{F}_b(c) = \left[\mathbf{F}_b^T(c) \mid \mathbf{F}_b^P(c) \right] \tag{2}$$

where $\mathbf{F}_a^T(c)$ and $\mathbf{F}_b^T(c)$ are the training sets for the two cameras, $\mathbf{F}_a^G(c)$ is the gallery set of camera a and $\mathbf{F}_b^P(c)$ is the probe set for camera b. The re-identification task is to rank all individuals in the gallery of known targets in terms of similarity to the probe.

In the following, for the clarity of exposition, we will omit in the notations the reference to the channel c.

3.1. Training KCCA

KCCA constructs the subspace that maximizes the correlation between pairs of variables. Feature mapping into a higher-dimensional space is performed by exploiting the kernel trick.

In our case, given corresponding feature vectors from a camera pair, for each channel, we denote $\mathbf{K}_a^{TT}$ and $\mathbf{K}_b^{TT}$ the kernel matrices of pairs from the training sets, $\mathbf{K}_a^{GT}$ the kernel matrix of pairs from the gallery and training sets, and $\mathbf{K}_b^{PT}$ the kernel matrix of pairs from training and probe sets.

The objective of KCCA is then to identify the projection weights $\boldsymbol{\alpha}, \boldsymbol{\beta}$ by solving:

$$\arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{\boldsymbol{\alpha}' \mathbf{K}_a^{TT} \mathbf{K}_b^{TT} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}' {\mathbf{K}_a^{TT}}^2 \boldsymbol{\alpha} \boldsymbol{\beta}' {\mathbf{K}_b^{TT}}^2 \boldsymbol{\beta}}}. \tag{3}$$

The norms of the projection vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are regularized in order to avoid trivial solutions according to [Hardoon et al. 2004].

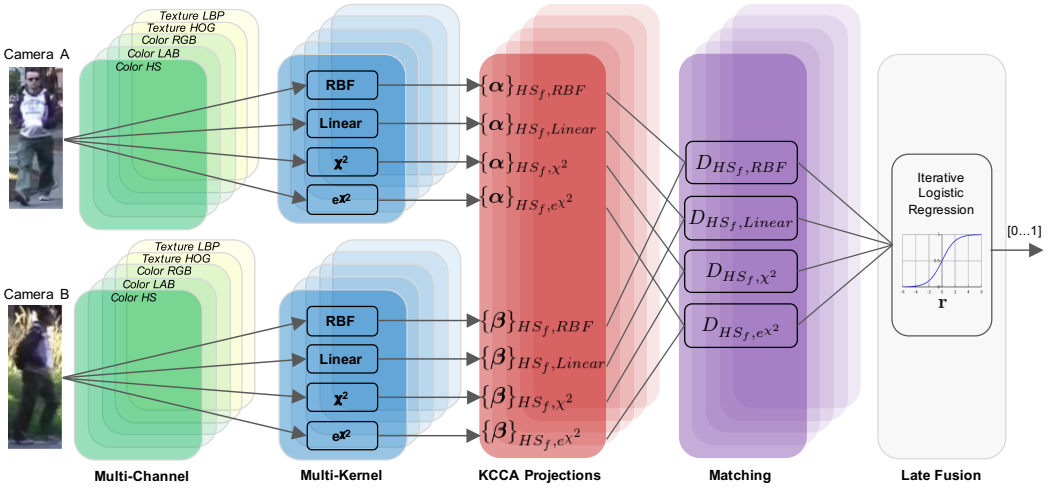


Fig. 2. An illustrative figure of our multi-channel, multi-kernel CCA (MCK-CCA) approach. Each feature channel is fed to different kernels: for the sake of clarity we show a single channel HS_f in the figure. For each of these combinations, we learn a specific KCCA projection and then use the learned projection to map each channel-kernel into its common subspace. Cosine distance is used to perform matching given a channel-kernel pair. Finally, distances coming from all the combinations are stacked together to form the distance vector. This distance vector is the input of an iterative logistic regression that performs the re-identification.

The top M eigenvectors of the standard eigenvalue problem obtained after regularization can then be applied as follows to project the gallery and probe data:

$$\tilde{\mathbf{F}}_a^G = \mathbf{K}^{GT}\boldsymbol{\alpha}\cdot\boldsymbol{\lambda} \quad (4)$$

$$\tilde{\mathbf{F}}_b^P = \mathbf{K}^{PT}\boldsymbol{\beta}\cdot\boldsymbol{\lambda} \quad (5)$$

where

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

are the learned projections and $\boldsymbol{\lambda}$ is the vector of eigenvalues obtained from KCCA. Weighting the learned projections with $\boldsymbol{\lambda}$ gives more relevance to those dimensions in the projected space that have higher eigenvalues, so improving the overall matching performance.

In order to improve re-identification, as a satisfactory common projection space for a channel and a camera pair may be obtained using a linear kernel or may require more complex kernel functions, we propose to learn multiple KCCA projections with four different kernels for each channel. Namely, we use a linear kernel, a Gaussian radial basis function kernel (RBF), a χ^2 kernel and an exponential χ^2 kernel and we denote $\mathcal{K}$ the set of kernels. We choose these kernels as they are commonly used but our method is not limited to those and can be easily extended to other kernels.

4. SELECTION OF THE OPTIMAL CHANNEL-KERNEL COMBINATIONS

According to our person representation, for each image pair, a distance vector of 80 values (four components with five features each, and four distinct kernels for KCCA) is defined. Our goal is hence to combine all the feature channels and kernels such that their combination results into the most effective re-identification. The overall process is represented in Fig. 2 and referred to as multi-channel, multi-kernel canonical correlation analysis (MCK-CCA).

In this section we detail how we formulate the matching process, how we weight each channel-kernel contribution, and how we select the best set of channel-kernel combinations.

4.1. Matching with logistic regression

We propose to formulate the matching probability of two samples using the logistic regression. Considering the feature vector $\tilde{\mathbf{f}}_{a_i}^G$ (from camera a) and the feature vector $\tilde{\mathbf{f}}_{b_j}^P$ (from camera b), we define as $\mathbf{d}_{ij}^{GP}$ the distance vector between these two feature vectors after KCCA projection concatenated with a *bias* term. The probability $\hat{p}(i,j)$ of these two features to represent the same person is calculated as:

$$\hat{p}(i,j) = \frac{1}{1+\exp(-\mathbf{r}'\mathbf{d}_{ij}^{GP})}, \tag{6}$$

where $\mathbf{r}$ is the weights vector.

4.2. Learning the logistic regression weights

Considering the training sets $\mathbf{F}_a^T$ and $\mathbf{F}_b^T$ from camera a and b respectively, the weights vector $\mathbf{r}$ in Eq. (6) are learned through the optimization of the logistic regression function:

$$\min_{\mathbf{r}} \frac{1}{2}\mathbf{r}'\mathbf{r} + C\sum_i \sum_j \log(1+\exp(-y_{ij}\mathbf{r}'\mathbf{d}_{ij}^{TT})) \tag{7}$$

where $\mathbf{d}_{ij}^{TT}$ is the distance vector between the feature vector $\tilde{\mathbf{f}}_{a_i}^T$ (sample i from camera a) and the feature vector $\tilde{\mathbf{f}}_{b_j}^T$ (sample j from camera b) after KCCA projection concatenated with a *bias* term; $y_{ij} = \{-1, 1\}$ accounts for the fact that the two features correspond to the same person in the two views, and C is a penalty parameter. The *bias* and the C parameter are selected using a two-fold cross-validation over the training sets $\mathbf{F}_a^T$ and $\mathbf{F}_b^T$, more details are given in Sect. 5.1. Note that the final model is trained over the whole training set using the best *bias* and C values from the cross-validation procedure.

4.3. Iterative learning of logistic regression weights

Positive weights indicate a non reliable distance obtained from a combination of a feature channel and a kernel. Let us denote $\mathbf{d}_{ij}^{GP}(c,k)$ the distance for one channel $c \in \mathcal{C}$ and one kernel $k \in \mathcal{K}$ and $\mathbf{r}(c,k)$ its corresponding weight in the logistic regression. The vector multiplication $\mathbf{r}'\mathbf{d}_{ij}^{GP}$ in Eq. (6) can be written as the sum of products $\mathbf{r}(c,k) \cdot \mathbf{d}_{ij}^{GP}(c,k)$ over all channels $c \in \mathcal{C}$ and kernels $k \in \mathcal{K}$. One can observe that a big distance value (that should correspond to a non matching pair) combined with a positive weight would actually lead to a lower denominator and thus a higher matching probability.

According to this observation, we derive an iterative filtering procedure to progressively drop out any channel-kernel that have positive weights. In particular, given a set of distances computed from the channel-kernels, we learn a logistic regression model. Channel-kernels that have a positive weights are removed and the remaining subset of distances is used to learn a new logistic regression model. This procedure is applied until there are no positive weights, and experimentally it never needed more than three iterations to reach that condition. That is, after at most three iterations all weights were negative, and thus the iterative procedure stopped.

5. EXPERIMENTS

We run our experiments on four standard publicly available datasets for re-identification that are VIPeR [Gray and Tao 2008], PRID [Hirzer et al. 2012b], PRID 450s [Roth et al. 2014] and CUHK01 [Li et al. 2013].

VIPeR [Gray and Tao 2008] presents illumination variations and pose changes between pairs of views. We split the whole set of 632 image pairs randomly into two sets of 316 image pairs, one for training and the other one for testing. The testing set is further split into a gallery and a probe set. A single image from the probe set is selected and matched with all the images from the gallery set. The process is repeated for all images of the probe set and the evaluation procedure is run on the 10 splits publicly available from [Farenzena et al. 2010].

The PRID dataset [Hirzer et al. 2012b] is generally considered being more challenging than VIPeR. It includes distractors as well as strong illumination changes across cameras. Differently from VIPeR, in this dataset the person images are acquired from above with similar poses. Camera view a contains 385 persons, camera view b contains 749 persons, 200 appearing in both views. These image pairs are randomly split into a training and a test set of equal size. For the evaluation, camera a is used as probe and camera b is used as gallery. Thus, each of the 100 persons in the probe set is searched in a gallery set of 649 persons (where 549 are distractors).

The PRID 450s [Roth et al. 2014] has almost the same characteristics of PRID but does not include distractors. Therefore, despite of the differences in appearance, the experimental setting for this dataset is similar to the one of VIPeR. This dataset contains 450 person image pairs captured by two cameras and image pairs are split in 225 for training and 225 for test.

The CUHK01 [Li et al. 2013] dataset, also known as CUHK Campus dataset, was captured with two cameras in a campus environment. Differently from the previous datasets, CUHK01 images have high resolution. It contains 971 persons, and each person has two images in each camera view. Persons are mostly captured in a frontal pose from camera a, and in a profile pose in the camera b with low illumination variations. The person identities are split into 485 for training and 486 for test. This datasets provides two evaluation modalities: single-shot, with one sample per subject (SvsS) and also multi-shot with, two samples per subject (MvsM N=2).

The evaluations for VIPeR, PRID and PRID 450s are conducted following a single-shot protocol. While on CUHK01 we perform both single- and multi-shot experiments. All the experiments are averaged over 10 trials.

5.1. Parameter settings

In our experiments, for the RBF and the exponential χ^2 kernels the normalization parameter σ has been estimated taking the median of all distances in the training set. As regards the KCCA, we set the reconstruction error of the Partial Gram-Schmidt Orthogonalization (PGSO) as $\eta = 1$ while we set the regularisation parameter as $\kappa = 0.5$, as in [Lisanti et al. 2014]. Finally, for the logistic regression we performed a two fold cross validation for both the penalty parameter C and the *bias* term to estimate their optimal values, in the range $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$ and $[1, \ldots, 500]$ respectively.

5.2. Comparison with techniques learning a common space

In Tab. I we report a comparison with the approaches that, similarly to our solution, learn a common space between two views in order to ease the re-identification problem. In this table we also highlight if a method learns a common space using deep learning (`DL`), if it uses a discriminative metric learning approach (`ML`) and if it uses non-linearity

Table I. Comparison with approaches learning a common subspace between views (the most similar to our approach). Techniques used in each approach (DL, deep learning; ML, metric learning; NL, non-linearity).

Dataset	Main Techniques			VIPeR			CHUK single-shot			CHUK multi-shot		
Rank	DL	ML	NL	1	10	20	1	10	20	1	10	20
Li *et al.* [Li and Wang 2013]	×	✓	×	29.6	69.3	85	–	–	–	–	–	–
ROCCA [An et al. 2015]	×	×	×	30.4	75.6	86.6	29.8	66.0	76.8	–	–	–
RCCA+Ref. Set [An et al. 2013]	×	×	×	30.3	74.7	86.8	30.0	67.8	77.0	–	–	–
RCCA+2 Ref. Set [An et al. 2016]	×	×	×	33.3	78.4	88.5	31.1	68.6	79.2	–	–	–
LOMO+XQDA [Liao et al. 2015]	×	✓	×	–	–	–	–	–	–	63.2	90	93
Siamese CNN [Ahmed et al. 2015]	✓	×	✓	34.8	75	–	47.5	80	–	–	–	
KCCA e^{χ^2} [Lisanti et al. 2014]	×	×	✓	36.8	84.5	92.3	38.1	74.2	82.4	47.7	84.3	90.8
MCK-CCA with filteredLR	×	×	✓	**47.2**	**87.3**	**94.7**	**57.0**	**86.8**	**92.2**	**69.5**	**93.6**	**96.3**

(NL). It is immediately clear that the use of non-linearity *per se* already matches the accuracy of other techniques that use reference sets [An et al. 2016; An et al. 2013] or Robust CCA (ROCCA) [An et al. 2015]. Non linearity is provided by the e^{χ^2} kernel in the work [Lisanti et al. 2014] or is obtained by training a deep CNN as in [Ahmed et al. 2015]. A single non-linear KCCA is more effective even than [Li and Wang 2013] which uses also metric learning. However, state-of-the-art performance among this type of methods is achieved by the interplay of different channels and kernels obtained with the proposed MCK-CCA, despite no metric learning is used in our approach. The proposed method indeed largely improves over the single kernel baseline and also over very recent methods [Liao et al. 2015; Ahmed et al. 2015].

5.3. Comparison with metric learning techniques

In this experiment we compare our strategy of learning common projection spaces against metric learning, both applied to our person representation. In particular, we compare the Large Margin Nearest Neighbor (LMNN) [Weinberger and Saul 2009] and the Logistic Discriminant-based Metric Learning (LDML) [Guillaumin et al. 2009] techniques with our multi-channel, multi-kernel CCA (MCK-CCA). The experimental setting is the following: in all the methods we employ our person representation composed of five features with four components; then, for each channel as defined in Sect. 2, we compute the LMNN, LDML and KCCA projections. We finally use the proposed iterative logistic regression to fuse all the distances together. This experiment is conducted on the VIPeR dataset and the performance is averaged over ten trials. We report the performance of our MCK-CCA obtained using only the linear kernel (MC-

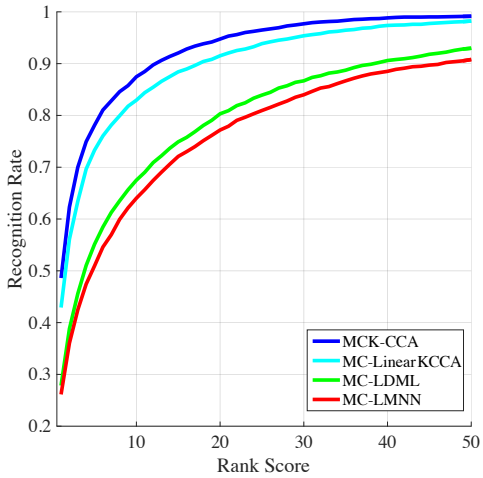


Fig. 3. Comparison of MCK-CCA with metric learning methods LMNN and LDML on the VIPeR dataset.

Table II. Comparative performance analysis at ranks {1,10,20,50,100} with respect to the state-of-the-art on VIPeR.

Dataset	VIPeR				
Rank	1	10	20	50	100
EIML [Hirzer et al. 2012a]	22	63	78	93	98
KRMCA [Liu et al. 2015a]	23.2	72.2	85.8	–	–
RPLM [Hirzer et al. 2012b]	27	69	83	95	99
eSDC [Rui Zhao 2013]	27	62	76	–	–
KLMM [Xiong et al. 2014]	28	76	88	–	–
Li *et al.* [Li and Wang 2013]	29.6	69.3	85	95	–
SalMatch [Zhao et al. 2013]	30.1	65	–	–	–
PLS+prototype [de Carvalho Prates and Schwartz 2015]	33	78	87	96	–
Siamese CNN [Ahmed et al. 2015]	34.8	75	–	–	–
CSL [Shen et al. 2015]	34.8	82.3	91.8	96.2	–
Improved DML [Yi et al. 2014b]	34.4	75.9	87.2	96.5	–
ECM [Liu et al. 2015b]	38.9	78.4	88.9	96.0	–
SLTRL [Wang et al. 2016]	39.6	78.3	87.9	–	–
LOMO+XQDA [Liao et al. 2015]	40.00	80.5	91.1	–	–
Semantic-Attribute [Shi et al. 2015]	41.6	86.2	95.1	–	–
mFilter [Zhao et al. 2014]	29.11	65	80	–	–
mFilter + LADF	43.4	82	95	–	–
Ensemble Metrics [Paisitkriangkrai et al. 2015]	45.9	**88.9**	**95.8**	**99.5**	**100**
KCCA e^{χ^2} [Lisanti et al. 2014]	36.8	84.5	92.3	98.6	99.8
MCK-CCA with LR	46.0	86.3	93.7	98.9	99.9
MCK-CCA with sparseLR	44.3	86.0	93.4	99.0	99.9
MCK-CCA with filteredLR	**47.2**	87.3	94.7	99.1	99.9

Linear KCCA) and all the kernels (MCK-CCA). In Fig. 3 we can see how MCK-CCA with the linear kernel only, improves over the two metric learning methods, under the same setting. The MCK-CCA with all the kernels achieves even higher performance. This experiment demonstrates that learning two projections, one for each camera, to map the data in a common space where features of the same person are highly correlated is more effective than learning a single metric [Weinberger and Saul 2009; Guillaumin et al. 2009]. Moreover, this observation is also supported by other recent methods which also propose the idea of learning both a metric and a common space to handle cross-view matching [Liao et al. 2015].

5.4. Comparison with the state-of-the-art

We now compare the performance of our approach with state-of-the-art methods. In particular we provide a side-by-side comparison of the proposed multi-channel, multi-kernel CCA (MCK-CCA) with recent state-of-the-art techniques such as: EIML [Hirzer et al. 2012a], RPLM [Hirzer et al. 2012b], eSDC [Rui Zhao 2013], SalMatch [Zhao et al. 2013], Li *et al.* [Li and Wang 2013], KLMM [Xiong et al. 2014], Improved DML [Yi et al. 2014b], mFilter [Zhao et al. 2014], PLS+prototype [de Carvalho Prates and Schwartz 2015], Siamese CNN [Ahmed et al. 2015], CSL [Shen et al. 2015], ECM [Liu et al. 2015b], LOMO [Liao et al. 2015], Ensemble Metrics [Paisitkriangkrai et al. 2015], SCNCD [Yang et al. 2014], KRMCA [Liu et al. 2015a], Semantic-Attribute [Shi et al. 2015], SLTRL [Wang et al. 2016]. For our method we both consider the case in which Logistic Regression with and without iterative filtering of the channel-kernels is used ("MCK-CCA with LR" and "MCK-CCA with filteredLR", respectively). Furthermore, we compare the performance of our proposed iterative filtering procedure with results obtained using a logistic regression model with a L1 constraint on the weights to enforce sparsity, that we dub "MCK-CCA with sparseLR". Results on all datasets show that our proposed iterative filtering method is a more effective way to select the best channel-kernels.

Table III. Comparative performance analysis at ranks {1,10,20,50,100} with respect to the state-of-the-art on PRID.

Dataset	PRID (with distractors)				
Rank	1	10	20	50	100
EIML [Hirzer et al. 2012a]	15	38	50	67	80
RPLM [Hirzer et al. 2012b]	15	42	54	70	80
Improved DML [Yi et al. 2014b]	17.9	45.9	55.4	71.4	–
Ensemble Metrics [Paisitkriangkrai et al. 2015]	17.9	50	62	–	–
KCCA e^{χ^2} [Lisanti et al. 2014]	16.6	49.3	61.0	78.6	89.9
MCK-CCA with LR	25.6	61.5	72.9	86.9	93.8
MCK-CCA with sparseLR	25.8	59.4	70.8	84.8	93.1
MCK-CCA with filteredLR	**26.9**	**62.3**	**73.6**	**87.9**	**94.2**

Table IV. Comparative performance analysis at ranks {1,10,20,50,100} with respect to the state-of-the-art on PRID 450s.

Dataset	PRID 450s				
Rank	1	10	20	50	100
SCNCD [Yang et al. 2014]	26.9	64.2	74.9	87.3	–
PLS+prototype [de Carvalho Prates and Schwartz 2015]	28	63	75	89	–
ECM [Liu et al. 2015b]	41.9	76.9	84.9	94.9	–
Semantic-Attribute [Shi et al. 2015]	44.9	77.5	86.7	–	–
CSL [Shen et al. 2015]	44.4	82.2	89.8	96.0	–
LOMO+XQDA [Liao et al. 2015]	58.2	90.1	**97.8**	–	–
SLTRL [Wang et al. 2016]	**59.4**	88.7	94.7	–	–
KCCA e^{χ^2} [Lisanti et al. 2014]	38.1	81.3	90.4	97.9	99.7
MCK-CCA with LR	55.5	90.4	95.2	98.5	99.9
MCK-CCA with sparseLR	55.6	90.3	94.8	**98.6**	99.9
MCK-CCA with filteredLR	55.6	**90.8**	95.4	**98.6**	**100.0**

In Tab. II we report the results on the VIPeR dataset. It is worth noticing that the Ensemble Metrics and our method, learning multiple metrics and projections respectively, to cope with the variations of pose and illumination of this dataset, score the best results. It appears that our MCK-CCA improves of few percentage points at rank-1 with respect to the Ensemble Metrics. The LOMO+XQDA [Liao et al. 2015] method exploits metric learning and performs feature projection into a common space between the two views as in our solution, although with a different method, but has a much lower performance. Finally, among the CNN-based methods the Siamese CNN [Ahmed et al. 2015] has the best performance but does not achieve a state-of-the-art result.

In Tab. III we show the recognition rate at various ranks on the PRID dataset. Our MCK-CCA outperforms all the other methods by a large margin. The use of multiple color features and multiple common projection spaces for each of them, permits dealing with the strong illumination differences between the views. All the solutions using a single representation appear to be less robust. Ensemble Metrics achieves 17.9% recognition rate at rank-1, less than our method by about 10%. It is worth to notice that our previous method [Lisanti et al. 2014] using a single feature and KCCA projection with a single kernel has comparable performance with Ensemble Metrics.

In Tab. IV we report the results on the PRID 450s dataset. On this dataset our MCK-CCA has a similar performance trend and comparable scores with SLTRL [Wang et al. 2016] and LOMO+XQDA [Liao et al. 2015]. Both methods aim at learning a transformation of the input to cope with appearance and pose variations. The SLTRL and LOMO+XQDA methods outperform our method at rank-1, but our method obtains the best performance at rank-10. This is likely to be caused by the fact that PRID 450s has appearance variations that challenge our person representation by making it less discriminative than LOMO: indeed, differently from other datasets, PRID 450s

Table V. Comparative performance analysis at ranks {1,10,20,50,100} with respect to the state-of-the-art on CUHK01 single-shot. (*) Method not directly comparable because it uses 100 subject for test and 871 for training.

Dataset	CUHK01 single-shot (SvsS)				
Rank	1	10	20	50	100
DeepReId (FPNN) [Li et al. 2014](*)	27.8	73	89	95	–
Semantic-Attribute [Shi et al. 2015]	31.5	65.8	77.6	–	–
Siamese CNN [Ahmed et al. 2015]	47.5	80	–	–	–
Ensemble Metrics [Paisitkriangkrai et al. 2015]	51.9	83.0	89.4	95.9	98.6
KCCA e^{χ^2} [Lisanti et al. 2014]	38.1	74.2	82.4	92.0	95.8
MCK-CCA with LR	55.9	86.1	91.7	96.4	98.4
MCK-CCA with sparseLR	55.8	86.0	91.5	96.3	98.6
MCK-CCA with filteredLR	**57.0**	**86.8**	**92.2**	**96.8**	**98.7**

Table VI. Comparative performance analysis at ranks {1,10,20,50,100} with respect to the state-of-the-art on CUHK01 multi-shot.

Dataset	CUHK01 multi-shot (MvsM N=2)				
Rank	1	10	20	50	100
mFilter [Zhao et al. 2014]	34.3	65	74	–	–
SLTRL [Wang et al. 2016]	61.6	90.2	94.4	–	99
LOMO+XQDA [Liao et al. 2015]	63.2	90	93	–	99
KCCA e^{χ^2} [Lisanti et al. 2014]	47.7	84.3	90.8	96.3	98.4
MCK-CCA with LR	65.7	91.9	95.8	98.7	**99.7**
MCK-CCA with sparseLR	65.5	91.5	95.9	**98.7**	99.7
MCK-CCA with filteredLR	**69.5**	**93.6**	**96.3**	98.5	99.6

presents unique characteristics such that pedestrians are observed from a top view and have slightly different scales between the two cameras.

On the CUHK01 dataset we performed comparisons for single-shot (SvsS) and multi-shot (MvsM N=2) modalities. Results are presented in Tab. V and Tab. VI, respectively. For both protocols the MCK-CCA outperforms the state-of-the-art by a significant margin. The capability of our MCK-CCA to correlate different representations into common projection spaces is very effective on this dataset.

5.5. Contribution of each channel-kernel

In Fig. 4 we show the contribution of each color space and texture feature employed in our person representation, separately for each kernel. For these experiments, the full person image is considered for feature extraction while upper, middle and lower regions are discarded to ease the analysis. The CUHK01 dataset has been used for this experiments since it is the largest dataset available among the ones used in our tests. The plots show the performance of each feature separately, the combination of all color spaces and the complete combination also including texture features. While it is evident that color histograms in the different color spaces contribute the most, nevertheless it is worth to note that the use of texture features allows obtaining even higher performance, e.g. from 30% to 40% at rank-1 for the linear and RBF kernels. This improvement is maintained across all the kernels we used. Finally, the combination of all channel-kernels in our iterative logistic regression pushes the recognition rate at rank-1 to 57%, as shown in Table V in Sect. 5.4.

In Fig. 5 we report the contribution of each component in our person representation for each color space histogram and texture feature, using a linear kernel, as well as their combination. For color features, it is possible to observe that each component performs differently. The full component outperforms the others in the 3 color spaces, nevertheless the combination of all components largely improves the performance. Texture features on the other hand are less effective for recognition, in particular when com-

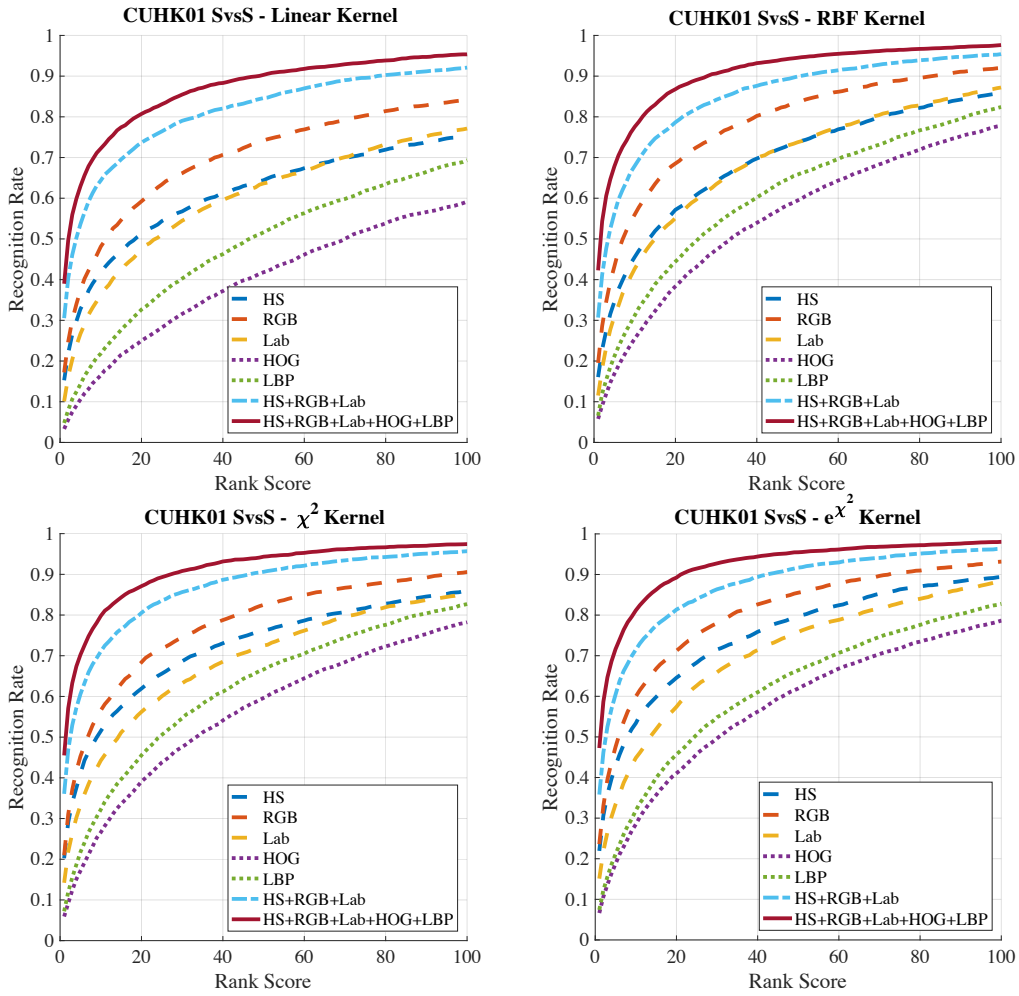


Fig. 4. Contribution of each color space histogram and texture feature, for the Linear, RBF, χ^2 and e^{χ^2} kernels, respectively, expressed in terms of CMC curves. Results are calculated considering only the feature extracted on the full person image.

puted on the lower region. This explains why our iterative logistic regression mostly drops this latter channel, as it will be shown in Sect. 5.6.

5.6. Analysis of iterative logistic regression

In this section, we analyze the performance of our iterative fusion scheme by giving insights on how each channel and kernel combination is filtered out by our iterative logistic regression. Additionally, we show the difference in performance between our MCK-CCA fusion scheme with respect to an early fusion KCCA baseline [Lisanti et al. 2014].

Each plot in Fig. 6 shows the probability of weight filtering per dataset: on the y-axes we report the channels, whereas on the x-axes we report the kernels. Moreover, for each figure, on the right part of the matrix, we show how many times a given channel is removed across all the kernels; instead, at the bottom, we show how many times a kernel is removed, across all the channels. We can see that MCK-CCA makes

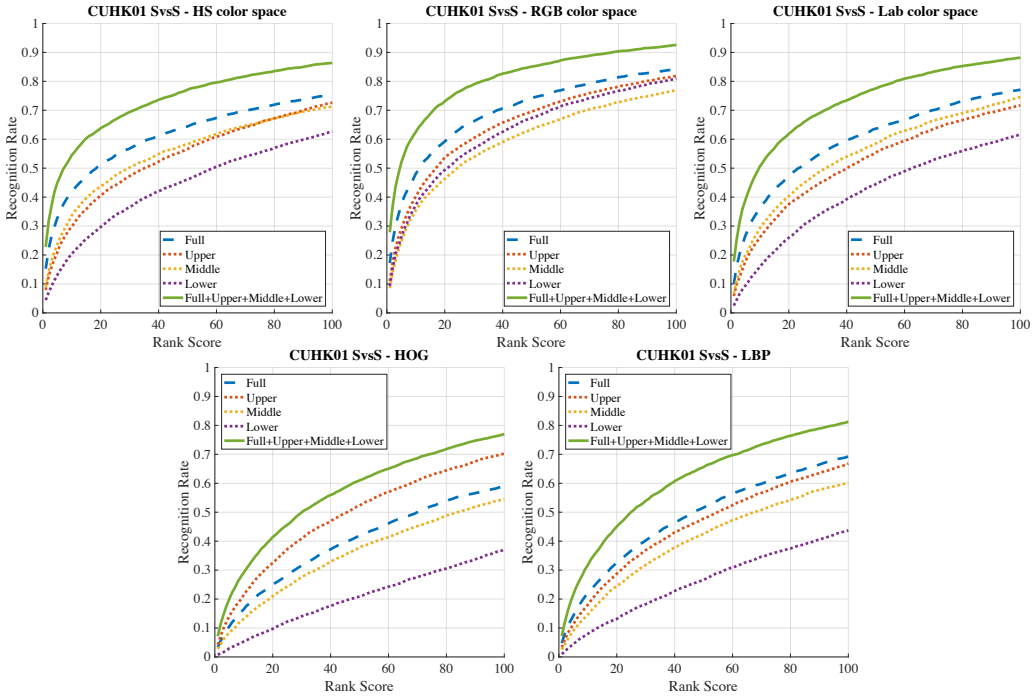


Fig. 5. Contribution of each component in our person representation for each color space (top) histogram and texture feature (bottom). Results are calculated considering only the linear kernel.

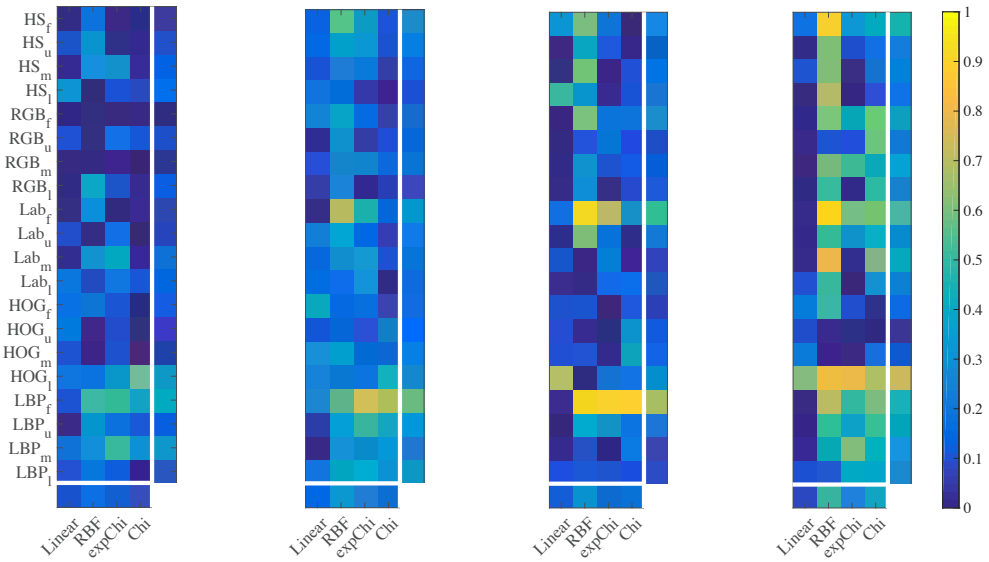
an extensive use of the iterative logistic regression filtering on all the datasets. VIPeR may be seen as an exception as most of the channel-kernels are often maintained. Our analysis on VIPeR shows that weak channels are represented by $\mathrm{HOG_l}$, $\mathrm{LBP_f}$ and $\mathrm{LBP_m}$. These channels correspond to texture features that may be noisy on VIPeR due to the low image resolution.

Regarding PRID, PRID 450s and CUHK01, we can observe that in general for texture features, the less relevant components are the full and lower region. Especially, the $\mathrm{LBP_f}$ and $\mathrm{HOG_l}$ are often filtered out. For color features, instead, these three datasets have in common that the channel $\mathrm{Lab_f}$ is removed with high probability, while is often maintained on VIPeR. It is also possible to see that, despite being largely used in literature, the RBF kernel is dropped out more frequently in general than the other kernels. This is reasonable since χ^2 kernels are usually better suited for histograms, which are used in our features.

Finally, considering all the results presented in Tables II, III, IV, V, VI, we can see that our late fusion outperforms, by a significant margin, a single KCCA learned over the stacked features, as proposed in [Lisanti et al. 2014]. This is mainly because a late fusion scheme allows maximizing the discriminative power of each channel-kernel combination. Moreover in most of the cases the iterative logistic regression scheme is able to select the most important channel-kernel combination and weight them in order to give more importance to the most discriminative ones.

5.7. Computational Complexity

Our approach makes extensive use of the kernel trick and multiple applications of KCCA. Even though this requires a computational effort at training time, the method remains still efficient with a moderate computational burden at test time. Moreover,



(a) VIPeR (b) PRID (c) PRID 450s (d) CUHK01

Fig. 6. Filtering analysis for each feature channel and kernel combination on the VIPeR, PRID 450s, PRID and CUHK01 datasets. Rows show channels, whereas columns show kernel. Each plot report also the summation over kernel (bottom) and the summation over channels (right).

thanks to the iterative logistic regressor, the method is able to discard some useless channel-kernel combinations to further speed-up the matching. Below, we separate the discussion on complexity regarding training and testing. These experiments are carried out on the CUHK01 dataset, as it is the largest dataset available among the ones that we used, providing a statistically significant test for the computational cost evaluation. Experiments are conducted on a workstation with 20 cores Intel Xeon@2.6GhZ and 256GB of RAM.

5.7.1. Computational effort at training time. The baseline complexity of KCCA depends on the number of training samples that are used. Assuming that we have N_T training samples to solve Eq. (3), then the complexity for solving a eigenvalue problem would be $\mathcal{O}(N_T^3)$. Although this complexity can seem prohibitive to be used for large amount of training samples, our method was able to perform the learning step of Sect. 3.1 seamlessly for the datasets we processed. Our approach learns offline 80 KCCA projections given by the combination of 20 channels times 4 kernels; the number of channels is given by 5 features times 4 components. Note that some of the KCCA projections that are learned are then discarded by the selection of the optimal channel-kernel combination (Sect. 4). This selection is then used at test time.

The timings for our learning part are shown in Fig 7(a) in function of the number of training samples $N_T = \{100, 200, 300, 400, 485\}$. Our approach can kernelize the training and gallery sets, learn the KCCA projections and estimate the logistic regression weights in 200s ($\sim$ 3 min) in the worst case. This also accounts for the time spent applying our iterative filtering procedure to select the most useful channel-kernels.

As a side note, our approach could still be applied even in those cases in which the re-identification application requires working at a large scale: approaches to learn KCCA at scale are based on random projections, a low-dimensional random feature space that approximates kernel evaluations [Rahimi and Recht 2007], or more recently, on

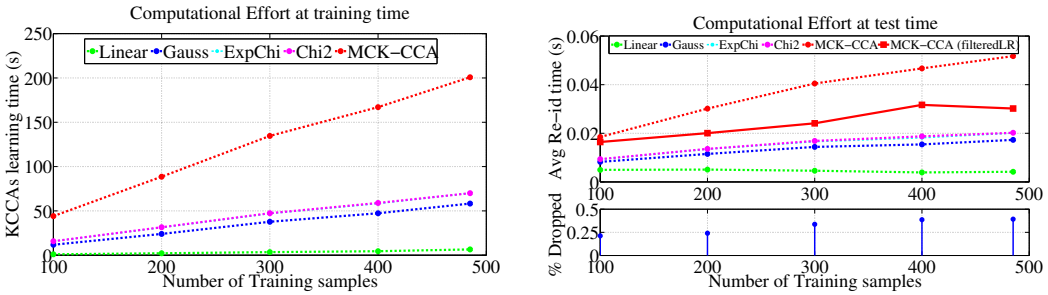


Fig. 7. *(a)* Computational effort at training time in function of the training size; *(b)* Average re-identification at test time in function of the number of training samples used. Overall computational effort using all channel-kernel combination (MCK-CCA) or dropping out some combinations of them (MCK-CCA filtered LR). In the bottom part: percentage of weights dropped by the iterative logistic regression in function of training samples.

the use of stochastic optimization in order to approximate KCCA [Wang and Livescu 2016].

5.7.2. Computational effort at test time. Much more important than the training time is the effort at test time. In this case the computational complexity of MCK-CCA remains quite efficient: the method needs to kernelize the probe with respect to the training set and then compute 80 linear projections using the learned KCCA basis; finally it performs normalized inner product comparison with the projected gallery. Moreover, the amount of KCCA projections used is reduced a priori by selecting the optimal channel-kernel combinations for each dataset. Fig. 7(b) shows the average re-identification time in seconds in function of the training samples used N_T. We can see the computational complexity for MCK-CCA when it uses all the channel-kernel combinations: the red-dashed curve shows how that does depend on the size of the training set used when we kernelize the probe. Moreover, we also broke down the complexity for each kernel and confirmed that the linear kernel is mostly invariant to the training samples. Interestingly, we can see how the proposed iterative logistic regression helps in speeding-up the performance: not only dropping logistic regression weights improves accuracy but decreases the computational effort of a noticeable amount. Fig. 7(b), at the bottom, reports also the percentage of weights that are dropped, when we apply our proposed method (MCK-CCA with filtered LR).

Considering these timings, our testing phase remains still applicable in real-time since we can perform re-identifcation in average at 0.03s (∼33Hz).

6. CONCLUSION

We have presented a method to overcome one of the main challenges of cross-view re-identification, that is dealing with drastic appearance changes. MCK-CCA grounds on the idea of addressing the extreme variability of person appearance in different camera views through multiple representations. These representations are projected onto multiple spaces that emphasize appearance correlation using KCCA and different kernels. Finally, our solution learns the most appropriate combinations for the observed pair through an iterative logistic regression, producing compelling results on standard re-identification benchmarks without impairing the computational complexity. The proposed technique showed also to be competitive with respect to state-of-the-art methods that learn a common subspace or use metric-learning. Investigating the pos-

sibility of directly incorporating metric-learning into our approach could represent an interesting line of research for future works.

REFERENCES

Ejaz Ahmed, Michael Jones, and Tim K. Marks. 2015. An Improved Deep Learning Architecture for Person Re-Identification. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 2, 10, 11, 12, 13

Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. 2013. Reference-based person re-identification. In *Proc. of the Int. Conf. on Advanced Video and Signal Based Surveillance*. 4, 10

Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. 2016. Person Reidentification With Reference Descriptor. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 4 (April 2016), 776–787. 4, 10

Le An, Songfan Yang, and Bir Bhanu. 2015. Person Re-Identification by Robust Canonical Correlation Analysis. *IEEE Signal Processing Letters* 22, 8 (August 2015), 1103–1107. 4, 10

Apurva Bedagkar-Gala and Shishir K Shah. 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing* 32, 4 (April 2014), 270–286. 2

Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. 2011. Custom Pictorial Structures for Re-identification. In *Proc. of the British Mach. Vision Conf.* 2

Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 5

Raphael Felipe de Carvalho Prates and William Robson Schwartz. 2015. Appearance-based person re-identification by intra-camera discriminative models and rank aggregation. In *Proc. of the Int. Conf. on Biometrics*. 3, 11, 12

Gianfranco Doretto, Thomas Sebastian, Peter Tu, and Jens Rittscher. 2011. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing* 2, 2 (January 2011), 127–151. 2

Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 2, 9

Douglas Gray and Hai Tao. 2008. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Proc. of the European Conf. on Computer Vision*. 9

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2009. Is that you? Metric learning approaches for face identification. In *Proc. of the Int. Conf. on Computer Vision*. 10, 11

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (December 2004), 2639–2664. 6

Martin Hirzer, Peter M. Roth, and Horst Bischof. 2012a. Person Re-identification by Efficient Impostor-Based Metric Learning. In *Proc. of the Int. Conf. on Advanced Video and Signal Based Surveillance*. 3, 11, 12

Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. 2012b. Relaxed pairwise learned metric for person re-identification. In *Proc. of the European Conf. on Computer Vision*. 9, 11, 12

Svebor Karaman and Andrew D Bagdanov. 2012. Identity inference: generalizing person re-identification scenarios. In *Proc. of the European Conf. on Computer Vision Workshops*. 4

Svebor Karaman, Giuseppe Lisanti, Andrew D. Bagdanov, and Alberto Del Bimbo. 2014. Leveraging local neighborhood topology for large scale person re-identification. *Pattern Recognition* 47, 12 (December 2014), 3767 – 3778. 4

Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2012. Large Scale Metric Learning from Equivalence Constraints. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 3, 4

Wei Li and Xiaogang Wang. 2013. Locally Aligned Feature Transforms across Views. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 4, 10, 11

Wei Li, Rui Zhao, and Xiaogang Wang. 2013. Human Reidentification with Transferred Metric Learning. In *Proc. of the Asian Conf. on Computer Vision*. 9

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 2, 13

Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 4, 10, 11, 12, 13

Giuseppe Lisanti, Iacopo Masi, Andrew D. Bagdanov, and Alberto Del Bimbo. 2015. Person Re-identification by Iterative Re-weighted Sparse Ranking. *Trans. Pattern Anal. Mach. Intell.* 37, 8 (August 2015), 1629–1642. 4, 6

Giuseppe Lisanti, Iacopo Masi, and Alberto Del Bimbo. 2014. Matching People across Camera Views using Kernel Canonical Correlation Analysis. In *Proc. of the ACM/IEEE Int. Conf. on Distributed Smart Cameras*. 4, 6, 9, 10, 11, 12, 13, 14, 15

Hao Liu, Meibin Qi, and Jianguo Jiang. 2015a. Kernelized relaxed margin components analysis for person re-identification. *IEEE Signal Processing Letters* 22, 7 (July 2015), 910–914. 3, 4, 11

Xiaokai Liu, Hongyu Wang, Yi Wu, Jimei Yang, and Ming-Hsuan Yang. 2015b. An Ensemble Color Model for Human Re-identification. In *Proc. of the Winter Conf. on App. of Computer Vision*. 3, 11, 12

T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Trans. Pattern Anal. Mach. Intell.* 24, 7 (July 2002), 971–987. 6

Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. 2015. Learning to Rank in Person Re-Identification With Metric Ensembles. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 3, 11, 12, 13

Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2010. Person Re-Identification by Support Vector Ranking. In *Proc. of the British Machine Vision Conference*. 4

Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Proc. of the Conf. on Neural Information Processing Systems*. 16

Peter M. Roth, Martin Hirzer, Martin Koestinger, Csaba Beleznai, and Horst Bischof. 2014. Mahalanobis Distance Learning for Person Re-Identification. In *Person Re-Identification*, Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen C. Loy (Eds.). Springer, London, United Kingdom, 247–267. 9

Xiaogang Wang Rui Zhao, Wanli Ouyang. 2013. Unsupervised Salience Learning for Person Re-identification. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 3, 11

Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. 2015. Person Re-Identification With Correspondence Structure Learning. In *Proc. of the Int. Conf. on Computer Vision*. 3, 11, 12

Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang. 2015. Transferring a Semantic Representation for Person Re-Identification and Search. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 3, 11, 12, 13

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 2

Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. 2013. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)* 46, 2 (November 2013), 29:1–29:37. 2

Jin Wang, Nong Sang, Zheng Wang, and Changxin Gao. 2016. Similarity Learning with Top-heavy Ranking Loss for Person Re-identification. *IEEE Signal Processing Letters* 23, 1 (January 2016), 84–88. 3, 11, 12, 13

Weiran Wang and Karen Livescu. 2016. Large-Scale Approximate Kernel Canonical Correlation Analysis. In *Proc. of the Int. Conf. on Learning Representations*. 17

Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research* 10 (June 2009), 207–244. 3, 10, 11

Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. 2014. Person Re-Identification Using Kernel-Based Metric Learning Methods. In *Proc. of the European Conf. on Computer Vision*. 3, 11

Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. 2014. Salient Color Names for Person Re-identification. In *Proc. of the European Conf. on Computer Vision*. 3, 11, 12

Dong Yi, Zhen Lei, and Stan Z. Li. 2014a. Deep Metric Learning for Person Re-Identification. In *Proc. of the Int. Conf. on Pattern Recognition*. 2

Dong Yi, Zhen Lei, and Stan Z. Li. 2014b. Deep Metric Learning for Practical Person Re-Identification. *arxiv preprint* abs/1407.4979 (2014). 2, 11, 12

Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Person Re-identification by Salience Matching. In *Proc. of the Int. Conf. on Computer Vision*. 3, 11

Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2014. Learning Mid-level Filters for Person Re-identification. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*. 3, 11, 13