Kernelized Saliency-based Person Re-Identification through Multiple Metric Learning

Niki Martinel* *Student Member, IEEE*, Christian Micheloni, *Member, IEEE*, and Gian Luca Foresti, *Senior Member, IEEE*

***Abstract*—Person re-identification in a non-overlapping multi-camera scenario is an open and interesting challenge. While the task can be hardly completed by machines, we, as humans, are inherently able to sample those relevant persons' details that allow us to correctly solve the problem in a fraction of a second. Thus, knowing where a human might fixate to recognize a person is of paramount interest for re-identification. Inspired by the human gazing capabilities, we want to identify the salient regions of a person appearance to tackle the problem. Towards this objective, we introduce the following main contributions. A kernelized graph-based approach is used to detect the salient regions of a person appearance, later used as a weighting tool in the feature extraction process. The proposed person representation combines visual features either considering or not the saliency. These are then exploited in a pairwise-based multiple metric learning framework. Finally, the non-Euclidean metrics that have been separately learned for each feature are fused to re-identify a person. The proposed KErnelized saliency-based Person re-identification through multiple metric LEaRning (KEPLER) has been evaluated on four publicly available benchmark datasets to show its superior performance over state-of-the-art approaches (e.g., it achieves a rank 1 correct recognition rate of 42.41% on the VIPeR dataset).**

***Index Terms*—Person Re-Identification, Kernelized Visual Saliency, Multiple Metric Learning, Dissimilarity Fusion**

I. Introduction

The person re-identification problem, i.e. identifying an individual moving across non-overlapping camera views, is receiving increasing attention from the community [1]. Many different applications, like situational awareness (e.g. [2], [3]), wide area scene analysis (e.g. [4], [5], [6]), etc. would benefit from it.

In spite of a swell of recent efforts, the re-identification is still an open issue due to a large number of hard challenges. Among all them, we can mention a few: (i) The time to move from one field-of-view (FoV) to another is not fixed and widely varies from person to person. Thus, putting temporal and spatial constraints is not feasible in this case. (ii) In a real scenario we are dealing with an uncontrolled environment where cameras are deployed with large FoVs, thus generating target images with low spatial resolution. This makes the acquisition of discriminating biometric features (e.g. face and gait features) hard as well as unreliable. Due to the poor quality of the acquired biometric features, methods relying on such features (e.g. [7], [8], [9]) perform unsatisfactorily. (iii) As a consequence, visual appearance features are still the first choice in re-identification problems (e.g. [1], [10], [11]). However, due to significant changes in viewing angle, lighting, background clutter, and occlusions, appearance features often undergoes large variations across non-overlapping camera views, hence, can be noticeably different from camera to camera.

Plenty of works have been devised to address the aforementioned visual appearance challenges: (i) *Discriminative signature based methods* (e.g. [10], [11], [12], [13]) exploit human-defined person signatures that are matched using distance measures like ℓ_2, χ^2, etc., or a combination of these. (ii) *Features transformation based methods* compute linear [14] and nonlinear [8], [15], [16], [17] transformation functions that are used to project features between different camera-dependent spaces. (iii) *Metric learning based algorithms* (e.g. [9], [18], [19], [20]) still rely on human-defined person signatures but model their spaces to learn non-Euclidean distances that are optimal for re-identification.

Despite such efforts, we believe that existing approaches lack three main aspects: (i) Most of the existing works compute the signatures either directly from the whole image or by fusing local features extracted from dense image patches. In such a process, each point of the person has the same importance. We believe that humans have a different approach and assign more importance to those particular points of a person that are useful for the re-identification. (ii) Assuming we can compute the importance of the points, it is not guaranteed that the same point is captured by all different camera views. (iii) Feature transformation functions are both highly non-linear [15], [16], [17] and depend on the class of the features, i.e., every feature transformation is modeled by a different function.

Our KErnelized saliency-based Person re-identification through multiple metric LEaRning (KEPLER) solution builds upon these limits and introduces three main contributions: (i) A new kernelized graph-based technique to compute the saliency (i.e., importance) of the points on the person. In such a scheme, a salient region is a consistent part of an image which is different from its surroundings and lies on the person silhouette. The computed saliency is used as a weight in the feature extraction process: the higher the saliency the higher the importance of the feature and vice versa. (ii) To handle occlusions, pose variations, etc. that make a same salient point not visible by two different camera views, saliency weighted features are supported by other ones that do not exploit it. (iii) A pairwise multiple metric learning framework is used to model each feature space separately rather than jointly.

The rest of the paper is organized as follows. In section II, an overview of the relevant work in the re-identification field is given. The proposed methodology is described in section III. In section IV, the superior performance of our method to existing ones are shown. Finally, conclusion are drawn in section V.

N. Martinel, C. Micheloni ang G.L. Foresti are with the Department of Mathematics and Comptuer Science, University of Udine, 33100, Italy.
E-mail: {niki.martinel, christian.micheloni, gianluca.foresti}@uniud.it
Manuscript received April 19, 2005; revised December 27, 2012.

II. Related Work

While there have been countless works in the field of tracking persons within camera FoVs (e.g. [21], [22], [23]), the re-identification problem is still in its infancy. Though many different categorization can be used to analyze the field [1], we group the existing literature into two main groups: (i) *biometrics-based* and (ii) *appearance-based* methods. Methods in such groups introduce different approaches, however, all of them aim to extract invariant features to build robust discriminating signatures and to use (or learn) proper distance measures that can be adopted to match a person across cameras.

In the following, we only introduce appearance based methods. A deep analysis of the field is out of the scope of this work, hence we redirect the interested reader to the surveys in [1], [7]. To clearly state the contribution of our work, we finally highlight the differences between our method and similar ones.

Appearance-based methods exploit appearance features by assuming that people do not change clothes as they walk between camera FoVs. Since the person re-identification problem can be viewed as an association problem where the goal is to track persons across camera FoVs, this is a reasonable assumption. As a matter of fact, clothes represent a feature that allows humans to recognize individuals [24]. Appearance-based methods can be further categorized into: (i) discriminative signature based methods, (ii) feature transformation based methods and (iii) metric learning based methods.

Discriminative signature based methods seek for highly distinctive representations to describe a person appearance under varying conditions. One of the early work following such an approach was proposed in [25]. A region-based segmented image was used to extract spatio-temporal local features from multiple consecutive frames. Local HSV-edgel features, Histogram of Oriented Gradients (HOG) and the spatial relationships between appearance labels were later exploited in [26]. Part-based clothing regions were used together with face features to build persons signatures [24], as well as to localize and match individuals in a 3D system determined by means of the structure-from-motion technique [27]. Dense grid patches were used to propose the Mean Riemannian Covariance Grid (MRCG) descriptor [28], later exploited in a boosting scheme [29]. Multiple local features [30], [31], also biologically-inspired [13], were used to compute discriminative signatures for each person using multiple images. In [10], re-identification was performed by matching shape descriptors of color distributions projected in the log-chromaticity space. Other methods adopted collaborative representations that best approximate the query frames [32], exploited reference sets to represent the whole body as an assembly of compositional and alternative parts [33] or use the similarity with the reference images as a new feature vector in a Regularized Canonical Correlation Analysis framework [34]. Recently, coupled dictionaries exploiting labeled and unlabeled data [35] and sparse discriminative classifiers ensuring that the best candidates are ranked at each iteration were proposed [11].

These methods addressed the problem by using human-defined representations that are both, distinctive and stable under changing conditions between different cameras. However, the exploited visual features are not invariant to the large variations that affect the images acquired by disjoint cameras.

Features transformation based methods have addressed the re-identification problem by modeling the transformation functions that affect the visual features acquired by disjoint cameras. In [15], a learned subspace of the computed brightness transfer function (BTF) between the appearance features was used to match persons across camera pairs. In [14], authors proposed to model the linear color variations between cameras using an incremental framework. Usually the modeled functions are used to transform the feature space of one camera to the feature space of another one. Then, the re-identification is performed in the so transformed feature space. Only recently, a few methods [36], [37], [38] had also considered the fact that the transformation is not unique and it depends on several factors (e.g. pose and viewpoint changes, image resolutions, photometric settings of cameras). In [39], a transfer learning framework was also introduced to deal with cases where target camera label information is not given.

Such methods have shown to be able to capture the transformation of features occurring between cameras, however, they still face problems when large intra-camera feature variations are present. The learning process used to capture such transformation is usually highly time consuming, hence not suitable for a real deployment.

Metric learning based algorithms lie in between the two aforementioned categories. Methods belonging to such a group still rely on particular features but also advantage of a training phase to learn non-Euclidean distances used to compute the match in a different feature space. In [40], a relaxation of the positivity constraint of the Mahalanobis metric was proposed. In [41], unfamiliar matches were given less importance in the optimization problem in a Large Margin Nearest Neighbor framework. In [42], multiple metrics specific to different candidate sets were learned in a transfer learning set up. In [18], the re-identification problem was formulated as a local distance comparison problem. Similarly, in [43] a learning framework was proposed to learn an optimal similarity measure. A distance metric from sparse pairwise similarity/dissimilarity constraints was introduced in [44]. In [45], a metric for biologically-inspired features and covariance descriptors was learned. In [20], a metric based on equivalence constraints was proposed. Such work has been extended in [46], where a smooth regularizer was introduced. In [19], regularized Local Fisher Discriminant Analysis was introduced to maximize the between-class separability and preserve multi-class modality. In [47], the re-identification in a camera network is formulated as a multi-task distance metric learning problem.

While learning a metric has shown to be promising for person re-identification, existing works assume that the optimal metric is suitable to match every feature and do not consider the fact that the joint feature space may be too complex to be correctly modeled.

Recently, *visual saliency based algorithms* have been investigated for re-identification purposes [48], [49]. In [48], [49], given the current image, the saliency is computed through a patch searching strategy with an image reference set. Differently from such works, we compute the image saliency just considering neighborhoods of pixels. This brings

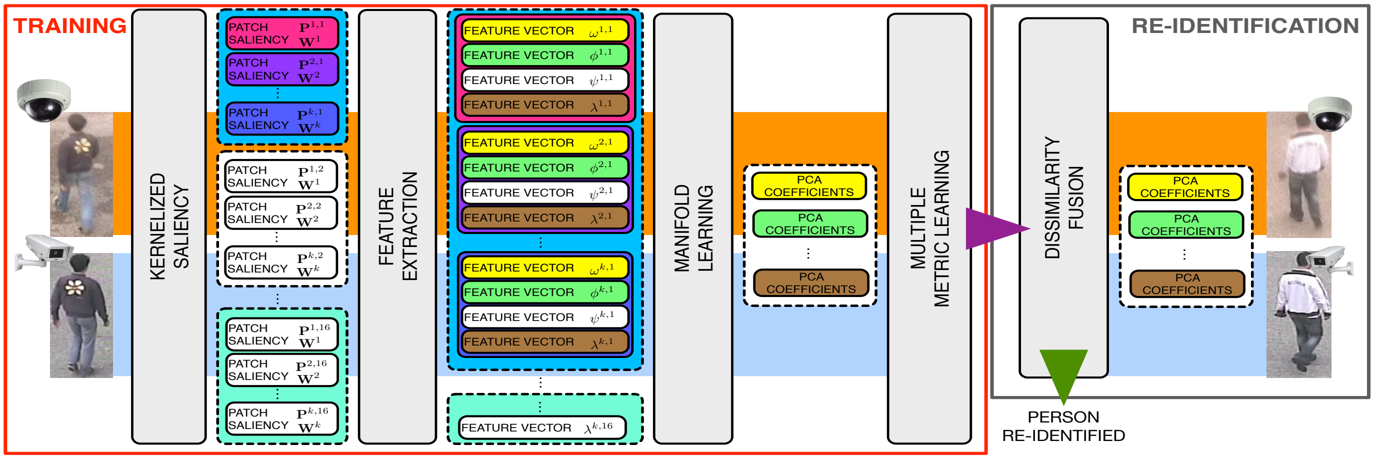


Fig. 1. Proposed system architecture based on five main stages: kernelized saliency computation, feature extraction, manifold learning, multiple metric learning and dissimilarity fusion. *(Best viewed in color)*

two main benefits: (i) lower computational requirements by avoiding the patch searching strategy adopted in [48], [49]; (ii) independence from a reference set that, as claimed in [48], [49], is robust as long as it well reflects the test scenario.

Differently from all such existing methods, in our approach we consider that: (i) points on each person have different importance; (ii) feature transformation depends on the class of the features, hence a single metric is not suitable to match different features.

III. The Approach

A. System Overview

As shown in Fig. 1, the proposed re-identification approach consists of five phases: (1) kernelized saliency computation, (2) feature extraction, (3) manifold learning, (4) multiple metric learning, and (5) dissimilarity fusion.

During training, each image is given to the kernelized saliency detection module (section III-B) that computes the saliency of each pixel, thus producing a saliency map. Both the saliency map and the image are split into overlapping patches which are then exploited by the feature extraction module (section III-C). Five different types of features are extracted from all the patches and for each color component of the selected color spaces. Features of the same type, extracted from the same color space, are concatenated and input to the manifold learning module that exploits Principal Component Analysis (PCA) to find the subspace where the extracted features lie. Finally, the whole training set of PCA reduced features is given to the multiple metric learning module (section III-D). This is in charge to learn a separate non-Euclidean metric between two cameras for each feature type and color space.

During the re-identification phase, the same reduced feature representations are computed for the two images acquired by the disjoint cameras. The obtained feature vectors, together with the learned metrics, are given to the dissimilarity fusion module. This computes the final dissimilarity which is finally used to tell if the two images are of the same person or not.

B. Kernelized Saliency

We usually tell that a portion of an image is "salient" if it is "different" from its surroundings. However, being our goal to re-identify a person moving between disjoint cameras we have to deal with background clutter that may induce state-of-the-art saliency detection algorithms [50], [51], [52] to label as "salient" a background region. We want only points on the person silhouette to have high saliency. On the basis of such considerations, we introduce a saliency computation approach that extends the algorithm in [53].

In [53], the following steps are adopted to compute the visual saliency: (i) Salient image points are detected by means of a Markov chain approach in which the transition probabilities are proportional to the features dissimilarity. (ii) Neighboring image points having high dissimilarity are grouped together using a Markov chain approach. (iii) The final saliency master map is computed as the weighted sum of the saliency maps obtained for the different features. Our Kernelized Graph-Based Visual Saliency (KGBVS) leverages these three steps and introduces the following contributions: (i) Different kernels are used in the computation of transition probabilities. The advantage of using them is twofold. First, using a kernel only neighboring points can lead to high saliency values. With this, the saliency has a more local meaning. In [53], the saliency is more global by considering also distant points that can naturally be very different. Second, the computed saliency can assume different meanings depending on the considered problem and used features. This can be achieved by properly selecting the kernels. Thus, the algorithm is more flexible. (ii) The saliency computation benefits from a visual saliency prior related to the person localization and shape.

Let $\mathbf{I} \in \mathbb{R}^{m \times n}$ be the image of a person and let assume that the silhouette stands somewhere in the center of it. Also, let $\mathbf{F} \in \mathbb{R}^{m \times n}$ be a feature map such that an element $\mathbf{F}_{x,y} = \pi(\mathbf{I}, x, y)$, where $\pi(\cdot)$ is a feature extraction function (e.g., wavelet transform, filter response, edge detector, etc.). Then, an activation map $\mathbf{A} \in \mathbb{R}^{m \times n}$ is computed such that an element $A_{x,y}$ has high value if (x, y) is in the center of the image and neighboring values of $\mathbf{F}_{x,y}$ are "different" one to each other. This is achieved as follows.

Let $G_{\mathbf{F}} = (V, E)$ be a fully-connected directed graph where $V = \{(x, y) | x = 1, \dots, m \wedge y = 1, \dots, n\}$ is the set of vertices. The weight of a directed edge $w((x, y), (p, q))_{\mathbf{F}} \in E$

is computed as

$$w((x,y),(p,q))_{\mathbf{F}} = \left|\log\left(\frac{\mathbf{F}_{x,y}}{\mathbf{F}_{p,q}}\right)\right| K_{\mathbf{F}}\left((x,y),(p,q)\right) \quad (1)$$

where the kernel function $K_{\mathbf{F}}$ returns values inversely proportional to the distance of the input points. The ratio between the two feature values represents the standard definition of dissimilarity. With respect to other measures it adapts better to the magnitude of the values. The absolute of the $\log$ allows to reach the lowest dissimilarity when the ratio is 1, while it returns higher values when the ratio is either lower or higher than 1. Once the graph is constructed, a Markov chain approach is exploited to detect the most dissimilar points of the image. For each node in V, its outbound edges weights are normalized to sum up to unity. These can be seen as the transition probabilities of a Markov chain. The equilibrium distribution computed on such a Markov chain effectively reveals the set of points that are most dissimilar from the others. Such a distribution defines the activation map $\mathbf{A}$.

When more feature maps are considered, different $\mathbf{A}$'s have to be fused. However, if the different activation maps $\mathbf{A}$ are uncorrelated, the additive fusion may lead to an uniform master map. To overcome such a problem, a new fully connected graph $G_{\mathbf{A}}$ is exploited to concentrate the mass of each $\mathbf{A}$'s into nodes with high activation values. The weight of the direct edge between two nodes (x,y) and (p,q) is computed as

$$w((x,y),(p,q))_{\mathbf{A}} = \mathbf{A}_{p,q} K_{\mathbf{A}}\left((x,y),(p,q)\right) \quad (2)$$

where $K_{\mathbf{A}}$ is a kernel function substituting the similarity measure in [53]. The outbound edges' weights of each node are normalized such that they sum up to unity. These, together with the set of vertex V, define another Markov chain. By finding its equilibrium distribution, represented by the concentrated activation map $\hat{\mathbf{A}}$, we have that most of the mass is accumulated around nodes of $\mathbf{A}$ having high activation values.

Let $\hat{\mathbf{A}}^j$ be the activation map computed for the j-th feature map $\mathbf{F}^j$, for $j = 1, 2, \dots, J$, then the final saliency master map is defined as

$$\boldsymbol{\Omega} = \mathcal{P}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{j=1}^{J} \boldsymbol{\alpha}_j \hat{\mathbf{A}}^j \quad (3)$$

where $\boldsymbol{\alpha}$ is a vector of weights and

$$\mathcal{P}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \sim \exp\left(-\left(\frac{x-\boldsymbol{\mu}_x}{\boldsymbol{\sigma_x}} + \frac{y-\boldsymbol{\mu}_y}{\boldsymbol{\sigma_y}}\right)\right) \quad (4)$$

is a non-isotropic Gaussian kernel "prior" centered at $\boldsymbol{\mu} = [\boldsymbol{\mu}_x, \boldsymbol{\mu}_y]^T$ with $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y]^T$, which accounts for silhouette location and shape. The saliency master map $\boldsymbol{\Omega} \in \mathbb{R}_+^{m \times n}$ is finally rescaled to $[0, 1]$ using the min-max normalization rule in [54].

The above formulation adds three important characteristics to [53]: (i) The kernels are used to control the weight of the two graphs edges. (ii) The kernels allow to achieve more flexibility. Different kernels might be used for different features, as well as for different classes of problems. For instance, the flexibility could be exploited when directional features are used. In such cases, related directional kernels could improve the saliency computation. (iii) The non-isotropic Gaussian kernel introduces a "bias" towards the image center. This means that lower mass is assigned to the nodes that will most probably belong to the background, which is compliant to the assumption that the person silhouette lies in the center of the image. A visual comparison of the achieved results is shown in Fig. 2.

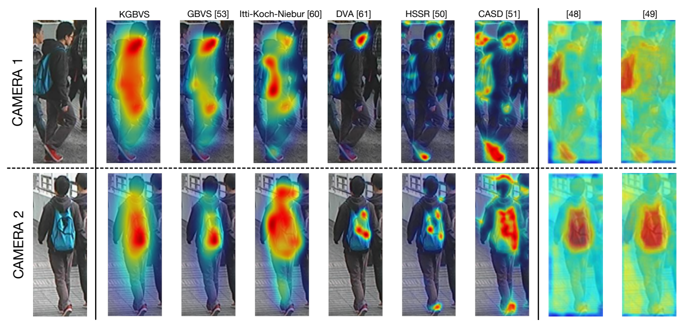


Fig. 2. Comparisons of the saliency detected by our method to state-of-the-art ones. The saliency gets higher as it goes from blue to red. First column shows the input images from two cameras. Second column is the saliency computed by the proposed approach. Third to seventh columns show the results achieved by existing methods that compute the saliency by considering neighborhoods of pixels. Last two columns show the results achieved by state-of-the-art methods that compute the saliency by means of a reference set. *(Best viewed in color)*

C. Feature Extraction and Manifold Learning

The proposed work wants to investigate the feasibility of using saliency to weight features for person representation. The idea is that image points that are salient are also discriminative, hence the features extracted from such points should be more important in the re-identification process. For such a task, there is a plethora of existing features that can be used. Some of these independently consider the information carried by a single pixel, while for others the information is given by the structure of groups of pixels (e.g. gradient, edges, etc.). Since the proposed saliency is pixel-based, the idea is to use it to weight the first category of features. This is a reasonable choice because the weighting procedure would only consider some pixels (i.e., the salient ones) more important than others. For each of the other features an ad-hoc weighting should be designed, which is not the main scope of this work.

Similarly to the majority of the existing approaches, color, shape and texture features are considered in the proposed work. Before extracting such features, the input image $\mathbf{I}$ is projected onto color space $S \in \{HSV, Lab, YUV, rgs^1, RGB, gray\}$. Then, the resulting image channels $\mathbf{I}^c$, $c = 1, \dots, 16$, and the saliency map $\boldsymbol{\Omega}$ are divided into k patches of equal size denoted $\mathbf{P}^{i,c}$ and $\mathbf{W}^i$ respectively, where $i = 1, \dots, k$ denotes the patch index.

For each patch i and channel c the following features are extracted: (a) the saliency weighted histogram ω computed as

$$\omega_{l,u}^{i,c} = \sum_{(x,y)\in\mathbf{P}^{i,c}} \begin{cases} \mathbf{W}_{x,y}^i & \text{if} \quad l < \mathbf{P}_{x,y}^{i,c} \leq u \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $\mathbf{W}_{x,y}^i$ and $\mathbf{P}_{x,y}^{i,c}$ are the saliency value and the pixel intensity at location (x,y) for patch i and color channel c.

[1] $r = R/(R+G+B)$, $g = G/(R+G+B)$, $s = (R+G+B)/3$

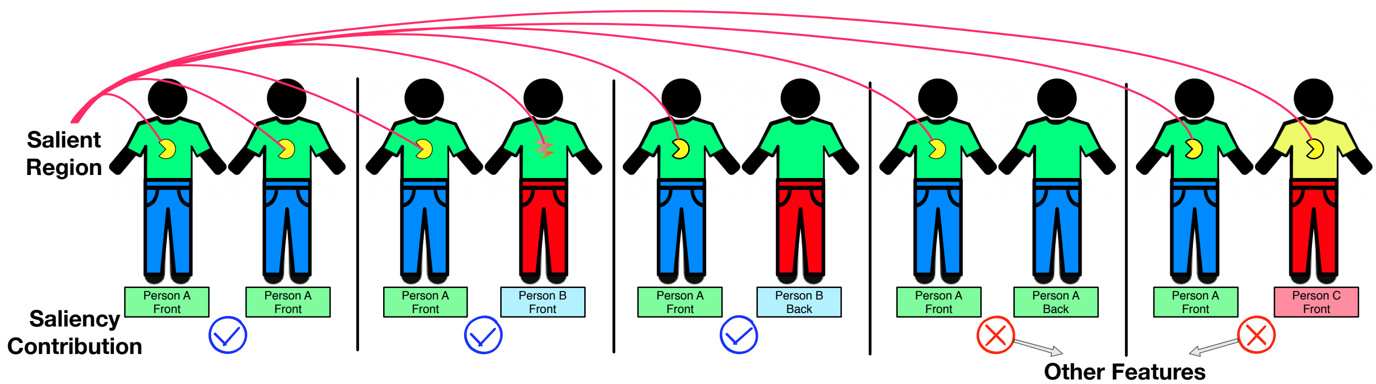


Fig. 3. Saliency weighted features are very discriminative for re-identifying a person based on the same salient region in the correct match (1^{st} example) or different region in the wrong match (2^{nd} and 3^{rd} examples). On the other hand, if due to different pose, orientation wrt. the camera, the correct match does not contain the salient region of the probe (4^{th} example) or if a wrong match does contain it (5^{th} example), then it is opportune to support the saliency weighted features with other features that do not consider the saliency. *(Best viewed in color)*

l and u are the lower and upper bin limits. (b) The color mean ϕ, (c) the 128-dimensional SIFT descriptor ψ, and (d) the Haar-like sparse-compressive features [55] λ. We also compute (e) the Local Binary Pattern (LBP) [56] γ from a grayscale representation of each patch i. Features of the same type extracted from all the k patches belonging to the same color space S are finally concatenated to get the corresponding feature vectors $\mathbf{x}_{(\omega,S)}, \mathbf{x}_{(\phi,S)}, \mathbf{x}_{(\psi,S)}, \mathbf{x}_{(\lambda,S)}, \mathbf{x}_{(\gamma,gray)}$. Notice that, while ω, ϕ, ψ and λ are extracted from all the five selected color spaces, γ is computed in the grayscale domain only.

In the current framework, the only feature that is pixel-based, hence exploits the saliency weighting mechanism is the histogram. It is well known that histograms do not have high discriminative properties (e.g., due to illumination and color changes, two different patches can generate a very similar histogram). We believe that discrimination can be introduced by means of saliency. Two different patches with similar histograms lying in different salient regions can be distinguished. In addition, as shown in Fig. 3, saliency weighted features can be very discriminative in different common cases, while for specific ones they require to be supported by other features that do not exploit saliency. This claim is substantiated by experimental results that show the importance of the saliency, but also how supporting it with other features strengthened the re-identification performance.

Due to the patch division the resulting feature vectors can be very high dimensional and each component may not have the same discriminative power. To address such an issue, supported by the studies in [57], we assume that the manifold where the extracted features lie is linear. Hence, we apply PCA to each feature vector separately to get the vector of coefficients $\hat{\mathbf{x}}_{(f,S)}$ where $f \in \{\omega,\phi,\psi,\lambda,\gamma\}$ denotes the feature type.

D. Multiple Metric Learning and Dissimilarity Fusion

For re-identification tasks, the input to metric learning algorithms is generally given by a vector representation of the image formed by joining multiple features (e.g. [18], [19], [20], [40]). Existing approaches have not considered that different types of features extracted from disjoint cameras may not be modeled by the same transformation function. The joint feature space may also be too complex to be robustly handled by a single metric. So, we propose to model each feature space separately. While any metric learning may be a suitable choice, since it has no parameters that need to be optimized, in this work we exploit the algorithm proposed in [20]. We briefly introduce it, then show how the learned metrics can be fused to compute the final distance.

The idea is to exploit statistical inference to find the optimal decision to establish whether a pair of features is dissimilar or not. This is achieved by setting the problem as a likelihood ratio test. Let $\mathbf{I}^A$ and $\mathbf{I}^B$ denote two images of persons A and B viewed by two disjoint cameras, and let h_0 be the hypothesis that A and B are not the same person ($(A,B)=0$) and h_1 the alternative one ($(A,B)=1$). By casting the problem in the space of pairwise differences $\hat{\mathbf{x}}^{A,B}_{(f,S)} = \hat{\mathbf{x}}^{A}_{(f,S)} - \hat{\mathbf{x}}^{B}_{(f,S)}$, the likelihood ratio can be defined as

$$\delta^{(A,B)}_{(f,S)} = \log\left(\frac{p\left(\hat{\mathbf{x}}^{A}_{(f,S)} - \hat{\mathbf{x}}^{B}_{(f,S)}|h_0\right)}{p\left(\hat{\mathbf{x}}^{A}_{(f,S)} - \hat{\mathbf{x}}^{B}_{(f,S)}|h_1\right)}\right). \tag{6}$$

Let suppose that the feature space of pairwise differences is governed by a normal distribution. Since $\hat{\mathbf{x}}^{A,B}_{(f,S)}$'s are symmetric we can assume the zero mean of the distribution, thus re-write the ratio test as

$$\delta^{(A,B)}_{(f,S)} = \log\left(\frac{\mathcal{N}\left(\hat{\mathbf{x}}^{A,B}_{(f,S)}, \mathbf{0}, \mathbf{\Sigma}_{(A,B)=0}\right)}{\mathcal{N}\left(\hat{\mathbf{x}}^{A,B}_{(f,S)}, \mathbf{0}, \mathbf{\Sigma}_{(A,B)=1}\right)}\right) \tag{7}$$

where $\mathbf{\Sigma}_{(A,B)=1}$ and $\mathbf{\Sigma}_{(A,B)=0}$ are the sum of outer products obtained by considering the pairwise feature differences $\hat{\mathbf{x}}^{A,B}_{(f,S)}$ computed for same or different persons, respectively.

By taking the log of eq.(7) and discarding the constant terms that provide an offset, we get

$$\delta^{(A,B)}_{(f,S)} = \left(\hat{\mathbf{x}}^{A,B}_{(f,S)}\right)^T \left(\mathbf{\Sigma}^{-1}_{(A,B)=1} - \mathbf{\Sigma}^{-1}_{(A,B)=0}\right)\left(\hat{\mathbf{x}}^{A,B}_{(f,S)}\right). \tag{8}$$

From eq.(8) we can learn the Mahalanobis metric $\mathbf{M}_{(f,S)}$ by clipping the spectrum of $\hat{\mathbf{M}}_{(f,S)} = (\mathbf{\Sigma}^{-1}_{(A,B)=1} - \mathbf{\Sigma}^{-1}_{(A,B)=0})$ computed through eigenanalysis. Then, the dissimilarity between the feature f extracted from images $\mathbf{I}^A$ and $\mathbf{I}^B$ projected

TABLE I
DETAILS AND COMPARISON OF COMMONLY USED PERSON RE-IDENTIFICATION BENCHMARK DATASETS.

Dataset	Persons	Image info	Cams	Additional Info
VIPeR [58]	632	Images: 1264 Avg. images per person per camera: 1 Size: 48×128	2	Scenario: outdoor Challenges: viewpoint variations, illumination changes and background clutter http://vision.soe.ucsc.edu/node/178
3DPeS [59]	191	Images:1012 Avg. images per person per camera: 3 Size: 31×100 to 176×267	8	Scenario: outdoor Challenges: viewpoint variations, not perfect detections, spatial resolution, illumination and color changes www.openvisor.org
CHUK02 (P1) [42]	971	Images:3884 Avg. images per person per camera: 2 Size: 60×160	2	Scenario: outdoor Challenges: viewpoint variations and illumination changes http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html
GRID [31]	1025	Images: 1275 Avg. images per person per camera: 1 Size: 29×67 to 181×384	8	Scenario: indoor Challenges: viewpoint variations, spatial resolution, color changes and image noise http://www.eecs.qmul.ac.uk/~ccloy/downloads_qmul_underground_reid.html

onto color space S is given by

$$d^2_{(f,S)}(\mathbf{I}^A,\mathbf{I}^B) = \sigma\left(\left(\hat{\mathbf{x}}^{A,B}_{(f,S)}\right)^T \mathbf{M}_{(f,S)} \left(\hat{\mathbf{x}}^{A,B}_{(f,S)}\right)\right) \quad (9)$$

where $\sigma(z) = \frac{1}{1+\exp^{-z}}$ ensures that $d^2_{(f,S)} \in [0,1]$.

Finally, the $d^2_{(f,S)}$'s computed using the learned Mahalanobis metrics can be fused to obtain the final dissimilarity between images of persons A and B as

$$D(\mathbf{I}^A,\mathbf{I}^B) = \sum_f \sum_S \boldsymbol{\beta}_{(f,S)} d^2_{(f,S)}(\mathbf{I}^A,\mathbf{I}^B) \quad (10)$$

where $\boldsymbol{\beta}_{(f,S)}$ is a vector of positive weights such that $\sum_f \sum_S \boldsymbol{\beta}_{(f,S)} = 1$, hence $D(\mathbf{I}^A,\mathbf{I}^B) \in [0,1]$.

Let $\mathcal{G}$ be the gallery set acquired by camera A (i.e., the set of persons for which labels are known) and $\mathcal{T}$ be the probe set acquired by camera B (i.e., the set of persons we want to re-identify), then, $\boldsymbol{\beta}$ is computed as follows:

$$\boldsymbol{\beta}_{(f,s)} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} R^i_{(f,S)} \quad (11)$$

where $i = 1, \ldots, |\mathcal{T}|$ denotes the i-th person in $\mathcal{T}$ and $R^i_{(f,S)}$ equals 1 if its true match has the lowest dissimilarity $d^2_{(f,S)}$ among all the gallery persons. Thus, $\boldsymbol{\beta}$ represents the re-identification performance achieved by each feature/color space. Feature/color spaces yielding to the highest rank 1 have more importance in the final dissimilarity fusion and vice versa. $\boldsymbol{\beta}$ is finally ℓ_1-normalized to satisfy $\sum_f \sum_S \boldsymbol{\beta}_{(f,S)} = 1$.

IV. EXPERIMENTAL RESULTS

We evaluated our approach on four publicly available benchmark datasets: the VIPeR dataset [58], the 3DPeS dataset [59], the CHUK02 dataset [42] and the GRID dataset [31]. We selected those on the basis of the following motivations: (i) the VIPeR dataset has strong illumination changes and viewpoint changes (most of the persons have viewpoint changes of about 90°); (ii) the 3DPeS dataset has images from 8 cameras. Persons are not always viewed by a frontal position and not perfect detections are present; (iii) the CHUK02 dataset has images of more than 900 persons appearing in two cameras. This is useful to see how the algorithm scales to a real scenario with lots of persons; (iv) the GRID dataset has more than 1000 persons, out of which, the majority is not present in all the cameras. Hence, such dataset resembles a real scenario in which we are not guaranteed that all the persons are viewed by all the cameras. Comparison and details of the datasets are given in Table I and reported in the following.

TABLE II
NUMBER OF RETAINED PRINCIPAL COMPONENTS AND VALUES OF ORIGINAL FEATURE DIMENSIONS (IN BRACKETS) FOR THE VIPER DATASET.

	ω	ϕ	ψ	λ	γ
HSV	49 (1080)	30 (1395)	35 (40320)	31 (900)	-
Lab	31 (1080)	31 (1395)	36 (40320)	30 (900)	-
YUV	54 (1080)	35 (1395)	45 (40320)	35 (900)	-
rgs	43 (1080)	27 (1395)	36 (40320)	30 (900)	-
RGB	38 (1080)	31 (1395)	28 (40320)	35 (900)	-
$gray$	-	-	-	-	54 (6195)

Evaluation Criteria: The re-identification mechanism commonly depends on how the gallery and the probe sets are organized. Let N be the number of images of each person in each of the two sets. Dependently on the value of N two matching philosophies are identified: i) single-shot ($N = 1$); ii) multiple-shot ($N > 1$). Two main approaches can be adopted to extend the single-shot to the multiple-shot case. Either we can take a statistic out of the $N \times N$ possible dissimilarities, or we can pool the features extracted from each of the N person images. While pooling seems to be a plausible operation (the average of all observations is likely to be an estimate of the centroid for all samples) it cannot handle the pose change of a person within a camera, e.g. if him/her moves straight, then turns. So, we have adopted a different approach. We have computed all the $N \times N$ dissimilarities and treated them as the probabilities of two persons A and B not being the same. Assuming these to be independent from each other, the joint probability has been obtained by multiplying all of them. Such a joint probability has been considered as the multiple-shot dissimilarity between the two persons A and B.

We report on the results for both a single-shot strategy and a multiple-shot strategy.

All the results are shown in terms of recognition rate by the Cumulative Matching Characteristic (CMC) curve and normalized Area Under Curve (nAUC) values. The CMC curve is a plot of the recognition performance versus the rank score and represents the expectation of finding the correct match within the top k ones. The nAUC describes how well a method performs irrespectively of the dataset size. For each dataset, the evaluation procedure has been repeated 10 times using independent random splits. We report on the average results

Fig. 4. 10 image pairs from the VIPeR dataset. The two rows show the different appearances of the same person viewed by two disjoint cameras.

computed for these 10 splits.

Implementation Details: In the adopted framework, we have considered the following settings. To compute and fuse the saliency maps of an image we have taken the same settings as in [53]. The Derrington Krauskopf Lennie color, intensity and orientation feature maps have been used. Both the kernel function $K_{\mathbf{F}}$ and $K_{\mathbf{A}}$ have been defined to be standard Radial Basis Functions with free parameter $\sigma = 1$. Each element of $\boldsymbol{\alpha}$ has been set to 1. The mean and standard deviation of the Gaussian kernel "prior" have been set to $\boldsymbol{\mu}_x = n/2$ and $\boldsymbol{\mu}_y = m/2$, and $\boldsymbol{\sigma}_x = n/4$ and $\boldsymbol{\sigma}_y = m/4$, respectively. We have sampled image patches of size 16×64 with a vertical stride of 8 pixels to extract the Haar-like and the weighted color histograms (each with 24 bins per channel). We have taken image patches of size 8×8 with a stride of 4×4 to compute the color mean. Similarly, LBP and SIFT features have been extracted from 50% overlapping patches of size 16×16. The dimension of the original feature vectors and the number of retained PCA coefficients are given in Table II. The number of PCA coefficients as well as all the aforementioned parameters have been selected by 5-fold cross-validation.

A. VIPeR Dataset

Due to the changes in illumination, low spatial resolution of images and viewpoint variations, the VIPeR dataset [58] is a tough person re-identification datasets. This dataset contains images of 632 persons viewed by two different cameras in an outdoor environment. Most of the image pairs have viewpoint changes larger than 90° (see Fig. 4). Since this dataset is considered the most challenging by the community, we provide a detailed performance analysis of our method on this dataset. For the evaluation, we have followed the common protocol as in [19], [49], [34] and resized all the images to 128×64. All the results provided in the following have been computed on the same 10 splits, using 316 person both for training and testing.

1) Features Performance Analysis: We have proposed to use different types of features extracted from images projected onto six different color spaces. To better understand how each feature/color space contributes to the re-identification, we have performed the following analysis.

Single feature analysis: First of all, we want to observe which of the considered features is the best performing one. Towards this objective, performances in Table III have been computed by independently extracting each feature from every color space. To verify if the adopted method to learn $\boldsymbol{\beta}$ correctly captures each feature importance, we have also reported the corresponding learned values.

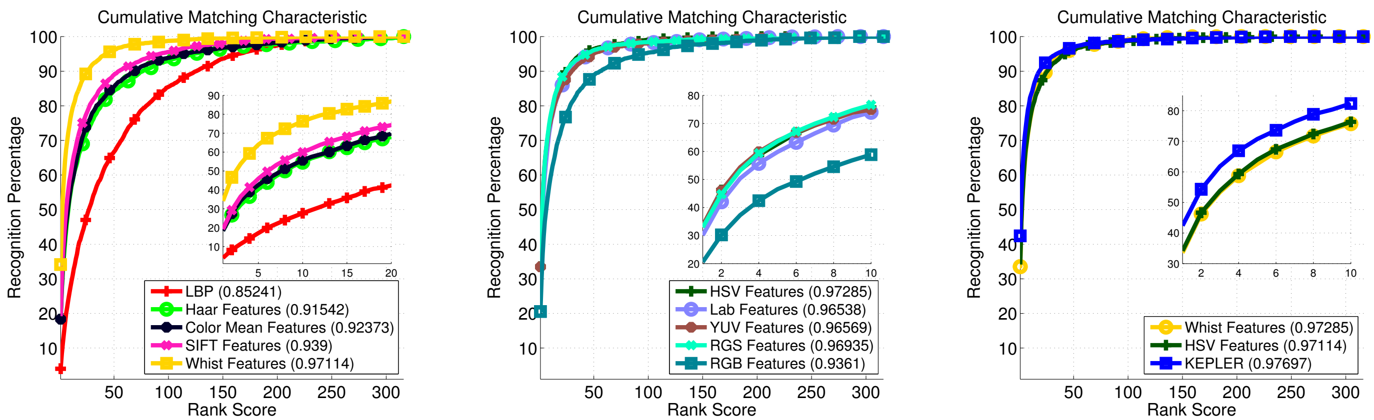
Results demonstrate that saliency weighted histogram features (ω) perform better than any other feature for every color space, except for the RGB one. Histograms extracted from the HSV color space obtain the highest rank 1 correct recognition rate (26.11%) and the best overall performance (with an nAUC value of 0.9645). The runner up is the saliency weighted histogram extracted from the RGS color space which has an nAUC value of 0.9593. All other features, apart from the color mean features extracted from the HSV color space, have a rank 1 recognition rate lower than 15%. LBP texture features (γ) have the lowest rank 1 recognition rate, i.e. 4.02% only. To show the benefits of the proposed saliency we also computed the performance using non-saliency weighted histogram features, denoted as $\widehat{\omega}$. Results show that for each color space, saliency weighted ones yield to better performance than those. In particular, on average, the rank 1 performances are improved by about 4% when saliency is used. Finally, results show that the learned $\boldsymbol{\beta}$ weights generally represent the test re-identification performance.

TABLE III
COMPARISON OF THE FEATURE PERFORMANCE ON THE VIPER DATASET. SECOND COLUMN SHOWS THE LEARNED $\boldsymbol{\beta}$ WEIGHTS FOR EACH FEATURE/COLOR SPACE, LAST 6 COLUMNS SHOW THE RECOGNITION PERFORMANCE FOR REPRESENTATIVE RANKS TOGETHER WITH THE NAUC VALUE. RESULTS FOR $\widehat{\omega}$ HAVE BEEN COMPUTED USING NON-SALIENCY WEIGHTED HISTOGRAM FEATURES. BEST RESULTS FOR EACH RANK ARE IN BOLDFACE FONT.

Rank →	$\boldsymbol{\beta}$	1	10	20	50	100	nAUC
HSV λ	0.0412	14.75	49.37	61.55	80.13	92.53	0.9076
Lab λ	0.0392	12.41	45.70	60.35	79.30	90.89	0.8941
YUV λ	0.0466	12.53	44.62	59.62	77.28	89.75	0.8880
RGS λ	0.0405	12.59	46.96	59.84	79.46	91.01	0.8967
RGB λ	0.0456	13.01	44.40	58.89	77.25	89.56	0.8871
HSV ϕ	0.0445	15.79	53.80	68.26	85.82	95.06	0.9268
Lab ϕ	0.0327	9.08	40.03	54.30	73.16	88.20	0.8727
YUV ϕ	0.0379	9.40	41.30	55.98	76.17	87.91	0.8810
RGS ϕ	0.0330	11.04	44.27	57.66	77.06	89.62	0.8871
RGB ϕ	0.0357	9.49	42.82	56.77	76.20	89.40	0.8855
HSV ψ	0.0371	8.35	39.15	54.91	77.75	90.41	0.8908
Lab ψ	0.0452	11.42	44.94	59.49	79.40	90.70	0.8990
YUV ψ	0.0628	10.73	44.11	59.75	80.63	91.71	0.9025
RGS ψ	0.0431	9.30	43.73	58.89	80.47	92.53	0.9036
RGB ψ	0.0160	3.58	21.17	34.40	57.15	76.74	0.7990
HSV ω	0.0833	**26.11**	**70.57**	**83.83**	**95.38**	**98.42**	**0.9645**
Lab ω	0.0506	20.00	63.70	77.85	91.61	97.34	0.9504
YUV ω	0.0894	25.09	65.38	77.63	91.71	97.47	0.9509
RGS ω	0.0746	25.38	69.11	82.47	93.58	97.79	0.9593
RGB ω	0.0450	11.30	44.53	60.00	80.98	92.91	0.9060
HSV $\widehat{\omega}$	0.0719	22.46	66.15	77.11	92.11	95.98	0.9421
Lab $\widehat{\omega}$	0.0451	17.21	59.46	74.24	88.64	94.32	0.9306
YUV $\widehat{\omega}$	0.0626	20.23	61.22	74.35	89.09	94.61	0.9340
RGS $\widehat{\omega}$	0.0649	20.98	64.66	79.02	90.23	94.77	0.9371
RGB $\widehat{\omega}$	0.0376	8.47	41.25	57.46	78.99	93.10	0.8983
Gray γ	0.0560	4.02	27.63	42.37	66.71	85.47	0.8523

Feature type analysis: Through the following analysis, we want to understand which feature type should be used to achieve the best performance. To support this, we have run the experiments considering each feature type separately. Given a feature type (e.g., SIFT, color mean, etc.), it has been extracted from every color space, then the corresponding dissimilarities have been fused using Eq.(10).

Results in Fig. 5a echo those in Table III, where the features achieving the highest performance are the histogram ones. However, by fusing the corresponding dissimilarities, rather than independently considering each one of them, a rank 1 recognition rate of 34.15% is obtained. This shows that, with respect to the results reached by using the same features



(a) (b) (c)

Fig. 5. Performance on the VIPeR dataset reported as CMC curves. The inside pictures show the performance on reduced rank ranges. In (a), results are computed by extracting each feature type from every color space, then fusing. In (b), results are computed by considering a particular color space from which all the features are extracted then fused. In (c), results computed by extracting all the features from every color space, then fusing, are compared to the best results shown in (a) and (b).

extracted from the HSV color space only, a performance improvement of about 8% is achieved. SIFT, color mean and Haar features perform similarly to each other but worse than histogram features. Indeed, they achieve a recognition rate of 19.21%, 18.35% and 18.48% for the same rank 1, respectively. LBP features are the worst performing.

Results show that histogram features outperform all the other ones. In addition, irrespectively of the considered feature, performance improves if the final distance is computed by fusing the dissimilarities between features of the same type, rather than considering a single feature only.

Color space analysis: We also want to identify the most suitable color space: given a particular color space, all the proposed features have been extracted, then the computed dissimilarities have been fused to get the final distance.

Results in Fig. 5b show that performances vary little between the HSV, CIELab, YUV and RGS color spaces. Indeed, the nAUC values computed for these color spaces differ from each other by less than 1%. The best overall performance as well as the highest rank 1 recognition rate (33.51%) is achieved when the HSV color space is considered. The worst performance is achieved using the RGB color space. In such a case, the rank 1 recognition rate is of 20.57% only.

Results demonstrate that similar performances are achieved by using one of the HSV, CIELab, YUV and RGS color spaces. Similarly to the feature type analysis, regardless of the exploited color space, better performance is achieved if features dissimilarities are fused to compute the final distance.

Overall analysis: Summarizing all the previous experiments, we expect that extracting all the features from all the color spaces and fusing them to get the final distance yields to the optimal performance. As shown in Fig 5c, the final solution (denoted as "KEPLER") clinch our argument by improving all the previous results. Rank 1 performance achieved by considering the saliency weighted histogram features extracted from every color spaces (see Fig 5a) is improved by more than 8%. Similarly, rank 1 performance increases by more than 9% with respect to the case when all the features are extracted from the HSV color space only (see Fig 5b).

As a conclusion, we can state that extracting all the features from all color spaces and fusing them yields to better performance than other previous solutions.

TABLE IV

COMPARISONS WITH STATE-OF-THE-ART SALIENCY-BASED RE-IDENTIFICATION ALGORITHMS ON THE VIPER DATASET. FIRST 7 ROWS SHOW THE RESULTS ACHIEVED BY PIXEL-BASED SALIENCY METHODS FOR RE-IDENTIFICATION -PERFORMANCES OF EXISTING SALIENCY METHODS USED WITHIN OUR RE-IDENTIFICATION PROTOCOL ARE GIVEN IN THE FIRST 5 ONES. LAST 4 ROWS SHOW THE RESULTS ACHIEVED BY STATE-OF-THE-ART RE-IDENTIFICATION APPROACHES THAT COMPUTE SALIENCY BY MEANS OF A REFERENCE SET. BEST RESULTS ARE IN BOLDFACE FONT.

Rank →	1	10	20	50	100	nAUC
GBVS [53]	38.11	80.95	90.01	96.34	98.64	0.9670
Itti-Koch-Niebur [60]	38.46	81.04	90.14	96.50	98.67	0.9684
DVA [61]	37.69	80.76	90.47	96.51	98.70	0.9694
HSSR [50]	38.79	80.95	90.49	96.56	98.72	0.9705
CASD [51]	39.14	81.14	90.44	96.60	98.77	0.9727
KEPLER(Only prior)	39.49	80.95	89.71	96.22	98.56	0.9634
KEPLER	**42.41**	**82.37**	**90.70**	**97.06**	**98.89**	**0.9770**
SalMatch [48]	30.16	65.54	79.15	91.49	98.10	0.9542
PatMatch [48]	26.90	62.34	75.63	90.51	97.47	0.9496
eSDC.ocsvm [49]	26.74	62.37	76.36	-	-	-
eSDC.knn [49]	26.31	58.86	72.77	-	-	-

2) Saliency and Multiple Metric Learning Contribution Analysis: We have proposed a method to compute the saliency by analyzing pixels neighborhoods of a single image and used it as a weighting tool in the feature extraction process. We have also introduced a method to learn multiple metrics (one for each extracted image feature) and fuse them to obtain the final distance. In the following, we analyze these two contributions to understand how much they add to the final goal.

Saliency analysis: To verify that the proposed KGBVS saliency method yields to better re-identification performance than state-of-the-art ones, we have studied the behavior of existing algorithms, namely GBVS [53], Itti-Koch-Niebur [60], DVA [61], HSSR [50] and CASD [51], within our re-identification protocol (first 5 rows in Table IV). Saliency has been computed by such methods, then the proposed feature extraction, manifold learning and multiple metric learning procedures have been exploited. In the last 4 rows of Table IV comparisons with existing re-identification approaches that use

TABLE V
MULTIPLE METRIC LEARNING AND SALIENCY RESULTS ON THE VIPER DATASET. BEST RESULTS ARE IN BOLDFACE FONT.

Rank →	1	10	20	50	100	nAUC
SML	14.59	51.84	66.08	83.04	93.42	0.9154
KGBVS + SML	20.23	63.86	77.84	92.14	97.51	0.9446
MML	39.12	80.16	89.63	96.09	98.02	0.9669
KEPLER	**42.41**	**82.37**	**90.70**	**97.06**	**98.89**	**0.9770**

saliency computed by means of a reference set are given.

Results in Table IV demonstrate that, with an nAUC value of 0.9727, CASD [51] yields to the best re-identification performance between existing saliency methods. Since CASD [51] has a salient definition similar to KGBVS, i.e., salient regions are dissimilar with respect to both their local and global surroundings, it is reasonable to claim that saliency detection algorithms designed for re-identification should consider both the local and global distinctiveness of a person appearance. Despite this, KEPLER outperforms the approaches by reaching the best overall performance and a rank 1 recognition percentage higher than 40%. If only the Gaussian prior is used, a rank 1 correct recognition rate higher than existing methods is achieved (39.49%). Results also show that the proposed method has the highest rank 1 score and the best overall performance among existing approaches that use a reference set to compute the saliency, namely SalMatch [48], PatMatch [48], eSDC.ocsvm [49] and eSDC.knn [49]. To conclude, by studying the behavior of existing saliency methods within our protocol we have shown that KGBVS yields to superior performance. Results also demonstrate that KEPLER outperforms saliency-based state-of-the-art re-identification methods. Thus, saliency computed by considering neighborhoods of pixels of a single image can be useful for re-identification purposes.

Multiple Metric Learning analysis: Through the following analysis we want to understand in which part the saliency and the multiple metric learning contribute to the final result.

To show this, we have conducted experiments by separately considering the KGBVS and the multiple metric learning (MML) components. When multiple metric learning is not used (SML), the extracted features have been concatenated, then PCA has been applied to reduce the dimension to 54 (such value has been found through 5-fold cross-validation). When the saliency weighting mechanism (KGBVS) is not used each entry in $\mathbf{\Omega}$ has been set to 1.

Let refer to Table V, in particular to the case where a single metric is learned (SML) and KGBVS is not used. Results show that, by exploiting the KGBVS method, rank 1 performance improves by 5%, while by using MML and no KGBVS, performance increases by 25%. Notice that SML and KGBVS+SML actually correspond to the performance achieved by using the proposed features without and with saliency, respectively, on KISSME [20]. By jointly using KGBVS and MML (i.e., KEPLER), the rank 1 recognition rate is improved by about 28%. Such results demonstrate that, while MML has a stronger impact on the performance, by jointly considering it with KGBVS the best result is achieved.

As a results of the previous analyses, we can draw the following conclusions: (i) better performance is achieved when the dissimilarities between all features extracted from all the color spaces are fused using the proposed weighted combination; (ii) both KGBVS and MML approaches should be jointly used to achieved the best re-identification results.

Fig. 6. Visual results on the VIPeR dataset. First column shows the probe image. Second column shows the top 20 matches. Last column shows the correct match. Correct matches are highlighted in green.

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE VIPER DATASET. BEST RESULTS ARE IN BOLDFACE FONT. (*) ONLY RESULTS REPORTED TO 2 ROUNDED DIGITS ARE AVAILABLE. (**) THE BEST RUN WAS REPORTED, WHICH CANNOT BE DIRECTLY COMPARED TO THE OTHER RESULTS.

Rank →	1	10	20	50	100	nAUC
KEPLER	**42.41**	**82.37**	**90.70**	**97.06**	98.89	**0.9770**
kBiCoV [45]	31.11	70.71	82.44	-	-	-
SalMatch [48]	30.16	65.54	79.15	91.49	98.10	0.9542
RCCA(*) [34]	30	75	87	96	**99**	0.9682
LAFT [36]	29.60	69.30	81.34	96.80	-	-
MtMCML [47]	28.83	75.82	88.51	-	-	-
MCE-KISS [62]	28.2	72.1	-	95.6	-	-
RPLM(*) [40]	27.34	69.02	82.69	94.56	98.54	0.9625
PatMatch [48]	26.90	62.34	75.63	90.51	97.47	0.9496
eSDC.ocsvm [49]	26.74	62.37	76.36	-	-	-
eSDC.knn [49]	26.31	58.86	72.77	-	-	-
WFS [38]	25.81	69.56	83.67	95.12	98.89	-
SSCDL [35]	25.6	68.1	83.6	-	-	-
RS-KISS [46]	24.5	66.6	81.7	93.5	98.0	-
LF [19]	24.18	67.12	81.38	94.12	-	-
CI(comb) [10]	23.15	58.11	69.45	86.53	-	-
eLDFV [63]	22.34	60.04	71.00	88.92	**99**	0.9447
IBML(*) [64]	22	63	78	93	98	0.9516
eBiCOV [13]	20.66	56.18	68.00	84.90	88.66	0.9105
KISSME [20]	19.60	62.20	74.92	91.80	98.00	0.9481
PCCA [44]	19.27	64.91	80.28	95.00	97.07	0.9536
PRDC [43]	15.66	53.86	70.09	87.79	92.84	-
KEPLER (best run)	**48.10**	**83.54**	**91.77**	**97.47**	**99.37**	**0.9795**
LMNN-R(**) [41]	20	68	80	93	99	0.9572

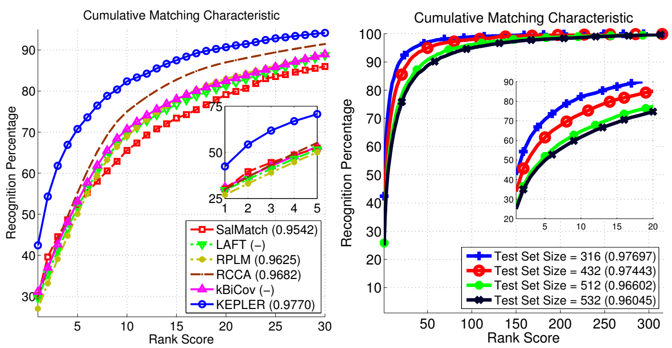
3) Comparison with State-of-the-art Methods: In the following the results of our KEPLER method are compared to the ones achieved by state-of-the-art approaches. We have considered the scenario where half of the dataset is used for training and the remaining half is used for re-identification [2].

As shown in Table VI, our method achieves the highest rank 1 recognition rate (42.41%), thus outperforming all existing approaches. It improves the previous top rank 1 performance [45] by more than 10%. A similar gap shows at ranks 10 and 20, where our method is the only one achieving a recognition percentage higher than 80% and 90%, respectively.

[2]Notice that some approaches are not using any training data as they are discriminative signature based methods (e.g. eBiCOV [13], etc.).

TABLE VII
COMPARISONS ON THE VIPER DATASET. RECOGNITION RATES PER RANK SCORE AS A FUNCTION OF THE TEST SET SIZE. BEST RESULTS ARE IN BOLDFACE FONT.

Test Set Size	432			512				532		
Rank →	1	10	20	1	5	10	20	1	10	20
KEPLER	**33.91**	**73.61**	**85.14**	**25.84**	**51.88**	**64.71**	**77.42**	**24.98**	**61.69**	**74.70**
RCCA [34]	22	59	75	-	-	-	-	15	47	60
MtMCML [47]	20	62	77	-	-	-	-	12	45	61
RPLM [40]	20	56	71	-	-	-	-	11	38	52
NRDV [65]	20	54	67	-	-	-	-	14	44	55
MCE-KISS [62]	14	49	69	-	-	-	-	-	-	-
RS-KISS [46]	10	40	61	-	-	-	-	-	-	-
PRDC [43]	13	44	60	9.12	24.19	34.40	48.55	9	34	49
MCC [43]	-	-	-	5.00	16.32	25.92	39.64	-	-	-
LAFT [36]	-	-	-	12.90	30.30	42.73	58.02	-	-	-
PCCA [44]	-	-	-	9.27	24.89	37.43	52.89	-	-	-



(a) (b)

Fig. 7. Results on the VIPeR dataset reported as averaged CMC curves. In (a), comparisons with state-of-the-art methods are shown. In (b), results are shown as a function of the test set size.

KEPLER also achieves the best overall performance with an nAUC value of 0.9770. Qualitative results are shown in Fig. 6.

In Fig. 7a, KEPLER performance are compared with five state-of-the-art approaches. For the considered ranks, our method outperforms existing ones. In particular, it achieves a recognition rate higher than 50% at rank 2 only. At rank 5 such a gap increases since KEPLER reaches a recognition rate of about 70% while all the other methods used for comparison achieve a recognition rate lower than 55%.

As commonly performed by state-of-the-art approaches (e.g. [34], [47], [40], [65]), we have run the experiments using different training/test set sizes (see Fig. 7b). This is done to study how the number of persons in the training set affects the performance (i.e., how many labeled image pairs are required to generalize well). We report on the performance using the 3 different splits introduced in [43]. Results show that the performances vary little on higher ranks but have differences on first ones. In particular, a rank 1 correct recognition rate of 33.91% is achieved when 200 persons are in the training set and the remaining 432 persons form the test set. This is a very interesting result if compared to the results reported in Table VI. Indeed, using less images to learn the proposed metrics, our method has better re-identification performance than the previously top performing approach [45].

In Table VII, we compare our method with exiting approaches using the same 3 splits. Results show that our method has the best performance on all the reported ranks for all the three considered partitions. In particular, for the case when 432 persons are in the test set, the proposed method outperforms the runner up by more than 11% at ranks 1 and 10. The same occurs when the test set contains 512 individuals. KEPLER outperforms all existing methods by more than 12% at rank 1 and by more than 20% at higher ones. Finally, our method achieves the best rank 1 recognition rate (24.98%) when 532 persons are considered as test set.

Results demonstrate that the proposed approach has superior performance than all existing ones on the considered dataset. In addition, the comparison analysis show that using KEPLER fewer samples are required to achieve good re-identification performance.

TABLE VIII
TIMING COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE VIPER DATASET.

Method	Rank 1	Appearance Modeling [sec]	Training Time [sec]
KEPLER (only prior)	39.49	3971	1.36
KEPLER	**42.41**	5299	1.37
RPLM [40]	27.34	3675	0.1
IBML [64]	22	3675	0.3
KISSME [64]	19.60	3524	**0.01**

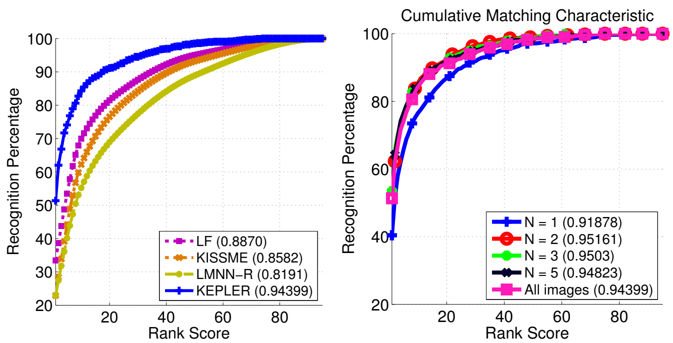
4) Computational Performance: In Table VIII we compare the computational times of our method and existing ones, namely RPLM [40], IBML [64] and KISSME [20], by using a MATLAB implementation executed on a 3.4 GHz Intel i7 CPU. The values show that the proposed appearance modeling requires more computational time. This is due to a greater number of used features with respect to the compared solutions. A similar trend is valid also for the training performance.

B. 3DPeS Dataset

The 3DPeS dataset [59] contains different sequences of 191 people taken from a multi-camera distributed surveillance system. Each of the 8 outdoor cameras is presented different light conditions and calibration parameters, so the persons were detected multiple times with different viewpoints. They were also captured at different time instants during the course of different days, in clear light and in shady areas. This results in a challenging dataset with strong variation of light conditions (see Fig. 8). The provided samples show that the 3DPeS dataset is composed by images including more persons or representing wrong detections. Moreover, like in typical video surveillance scenarios, the angle between the optical axis and the vertical axis of a person can vary noticeably from camera to camera.

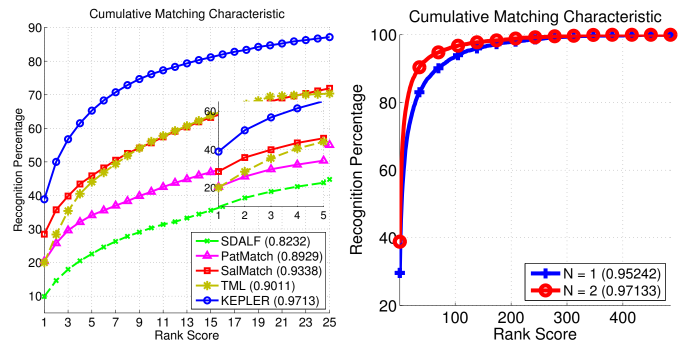
Fig. 8. 10 image pairs from the 3DPeS dataset. The two rows show the different appearances of the same person viewed by two disjoint cameras.

Fig. 10. 10 image pairs from the CUHK02 dataset. The two rows show the different appearances of the same person viewed by the two disjoint cameras.



(a) (b)

Fig. 9. Results on the 3DPeS dataset reported as CMC curves. In (a), we compare our results to state-of-the-art methods: LF [19], KISSME [20] and LMNN-R [41]. In (b), performances are shown as a function of the number of shots used during both the training and the re-identification phase.



(a) (b)

Fig. 11. Results on the CUHK02 Campus dataset (Camera P1) reported as averaged CMC curves. In (a), we show our superior performance to state-of-the-art approaches: SDALF [30], TML [42], PatMatch [48] and SalMatch [48]. In (b), results of the proposed method are given as a function of the number of images per person used for training and testing.

We compare our results to the ones reported in [19]. However, as in [19] no much details were given about how the results had been computed, we follow a similar approach to the one used in the VIPeR dataset and resize all the images to 128×64 pixels. As this dataset comes with more than a single image per person per camera, we have considered that all images were used to compute the results in [19]. Then, as in [19], we have randomly split the dataset into a training set and a test set containing 95 persons each.

TABLE IX
COMPARISON OF THE PROPOSED METHOD ON THE 3DPES DATASET. BEST RESULTS ARE IN BOLDFACE FONT.

Rank →	1	10	25	50	nAUC
KEPLER ($N = 1$)	40.42	76.53	89.79	97.16	0.9188
KEPLER (All images)	**51.37**	**84.32**	**92.63**	**98.53**	**0.9440**
LF [19]	33.43	69.98	84.80	95.07	0.8870
KISSME [20]	22.94	62.21	80.74	93.21	0.8582
LMNN-R [41]	23.03	55.23	73.44	88.92	0.8191

In Fig. 9a comparisons with state-of-the-art approaches, namely LF [19], KISSME [20] and LMNN-R [41] are shown. Our method achieves better performance than all existing ones at every considered rank. In particular, as shown in Table IX, at rank 1, KEPLER achieves a correct recognition rate of 51.37% while, LF [19], KISSME [20] and LMNN-R [41] achieve a recognition rate of 33.43%, 22.94% and 23.03%, respectively.

In Fig. 9b performances of our method are shown as a function of N. Since not all the persons come with an equal number of images, if the selected value of N was higher than the actual number of available images, the maximum allowable number of images for that person has been taken. When the single shot approach is considered, our method achieves a recognition percentage of 40.42% at rank 1 and a recognition percentage of 89.79% when the considered rank is 25. Considering a multiple-shot modality, the performances remain consistent either using $N \in \{2, 3, 5\}$ or all the available images. This is confirmed by the fact that the reported nAUC values change by less than 1% among all the three cases.

C. CUHK02 Campus Dataset

The CUHK02 Campus dataset [42] has images acquired by disjoint camera views in a campus environment. The dataset comes with 1,816 persons and five camera pairs denoted P1–P5 each of which is composed by different sensors (i.e. the dataset has images from ten camera views). The five camera pairs have 971, 306, 107, 193 and 239 persons, respectively. Each person has two images in each camera. Other than being challenging for pose variations, this dataset is the one that has the highest number of persons collected by a single camera pair. To evaluate our method and compare it to the state-of-the-art we have followed the same protocol used in [48], [42]. Results are reported for camera pair P1 when $N \in \{1, 2\}$ is considered. In this camera pair, images from the first camera are captured from lateral view, while images from the second camera are acquired from a frontal view or back view (see Fig. 10). All the 3,884 images have been resized to 160×60. The dataset has been split into a training set containing 485 pedestrians and a test set having images for the remaining 486.

In Fig. 11a, we compare the results of our method to four state-of-the-art approaches, namely, SDALF [30], PatMatch [48], SalMatch [48] and TML (Our_Generic) [42]. In the reported results, $N = 2$ images have been used both to learn the metric and to re-identify the targets. At rank 1 our method performs better than all other ones by reaching a correct recognition rate of 38.85%, thus improving the performance of SalMatch [48] by about 9%. As the rank score

TABLE X
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE CUHK02 (P1) DATASET. BEST RESULTS ARE IN BOLDFACE FONT.

Rank →	1	5	10	20	50	100	200	nAUC
KEPLER ($N = 1$)	29.59	52.95	64.12	74.76	86.97	93.55	97.86	0.9524
KEPLER ($N = 2$)	**38.85**	**65.28**	**76.12**	**84.72**	**92.85**	**96.58**	**98.72**	**0.9713**
SalMatch [48]	28.45	45.85	55.67	67.95	84.52	92.26	96.08	0.9374
PatMatch [48]	20.39	34.12	41.09	51.56	72.46	87.91	94.73	0.9065
TML(Our_Generic) [42]	20.53	45.54	56.61	69.62	85.74	93.75	-	-
SDALF [30]	9.90	22.57	30.33	41.03	55.99	67.39	84.12	0.8684

Fig. 12. 10 image pairs from the GRID dataset. The two rows show the different appearances of the same person viewed by two disjoint cameras.

TABLE XI
COMPARISON OF THE PROPOSED METHOD ON THE GRID DATASET. BEST RESULTS ARE IN BOLDFACE FONT.

Rank →	1	5	10	15	20
KEPLER	**18.40**	**39.12**	**50.24**	**57.04**	**61.44**
PRDC [43]	9.68	22.00	32.96	38.96	44.32
PRSVM [67]	10.24	24.56	33.28	39.44	43.68
MRank-PRDC [66]	11.12	26.08	35.76	41.76	46.56
MRank-RankSVM [66]	12.24	27.84	36.32	42.24	46.56
MtMCML [47]	14.08	34.64	45.84	52.88	59.84

increases the gap with the runner up is more evident, resulting in an 18% averaged over ranks 5 to 20 (see Table X).

In Fig. 11b, the performances of the proposed method are shown as a function of N. As for the other datasets, by increasing the number of images used to learn the proposed metrics results improve. In particular, using $N = 1$ images the nAUC value is 0.9524, while with $N = 2$ it reaches 0.9713.

D. GRID Dataset

The QMUL underGround Re-IDentification dataset (GRID) [31] has images acquired by 8 disjoint camera views installed in a busy underground station. Under this scenario a sample of 1275 images of 1025 individuals has been taken to build the dataset. Out of the 1025 persons, only 250 appear in all camera views. Apart from the high number of persons, the dataset is challenging due to variations of pose, colors, lighting changes; as well as poor image quality caused by low spatial resolution (see Fig. 12 for a few examples). To evaluate our method and compare it to the state-of-the-art, we have followed the protocol in [66]. The dataset has been split into a training set and a test set each of which contains 125 pedestrians that are viewed in all the cameras. In the test phase, the 125 persons appearing in all camera views are selected as probes. The 125 corresponding matching persons, plus the remaining 775 non-paired persons form the gallery set. Hence, for each of the 125 probes there are 900 gallery persons to match.for each of the 125 probes

In Table XI we compare the results of our method to 4 state-of-the-art ones, namely PRDC [43], PRSVM [67], MRank-PRDC [66], MRank-RankSVM [66] and MtMCML [47]. As shown, our method outperforms all the existing ones at all the considered ranks. In particular, the previous top rank 1 performance is increased by more than 4%. The same occurs for higher ranks where our method is the only one that achieves a rank score higher than 50% and 60% at ranks 10 and 20, respectively.

V. CONCLUSION AND FUTURE WORK

In this work, we have proposed to address the re-identification problem by introducing a novel algorithm able to identify the salient regions of a person. A kernelized saliency approach giving high weights to the regions that are in the center of the image has been designed for such a purpose. The computed saliency is used as a weight in the feature extraction process which also combines other features that do not consider it. The manifold where the extracted features lie is learned through PCA, and the resulting coefficients are input to the proposed pairwise-based multiple metric learning framework. The obtained metrics are exploited to learn the coefficients of a linear combination used to compute the dissimilarity between image pairs. The superior performance of the proposed method to state-of-the-art ones have been shown through extensive evaluations conducted on four challenging benchmark datasets. Results have shown that all previous top rank 1 scores have been outperformed, as well as, less manually annotated data is needed to meet and surpass state-of-the-art re-identification performance. In particular, less data is required both in terms of number of persons in each camera, and in terms of number of images per person.

Finally, since the reported performance are promising and supporting the usage of saliency for feature weighting, we are considering for future work to study how saliency can be used to weight patches or any kind of feature.

REFERENCES

[1] R. Vezzani, D. Baltieri, and R. Cucchiara, "People Re-identification in Surveillance and Forensics: a Survey," *ACM Computing Surveys*, vol. 46, no. 2, 2014.

[2] C. Alcaraz and J. Lopez, "Wide-Area Situational Awareness for Critical Infrastructure Protection," *IEEE Computer*, vol. 46, no. 4, pp. 30–37, 2013.

[3] J. C. San Miguel, C. Micheloni, K. Shoop, G. Foresti, and A. Cavallaro, "Self-Reconfigurable Smart Camera Networks," *IEEE Computer*, vol. 47, no. 5, pp. 67–73, 2014.

[4] N. M. Nayak, Y. Zhu, and A. K. Roy-chowdhury, "Exploiting Spatio-Temporal Scene Structure for Wide-Area Activity Analysis in Unconstrained Environments," *IEEE TIFS*, no. 99, pp. 1–1, 2013.

[5] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems." *IEEE TPAMI*, vol. 34, no. 10, pp. 2064–70, 2012.

[6] E. Ricci, G. Zen, N. Sebe, and S. Messelodi, "A Prototype Learning Framework using EMD: Application to Complex Scenes Analysis." *IEEE TPAMI*, vol. 35, no. 3, 2012.

[7] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[8] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis." *IEEE TPAMI*, vol. 27, no. 12, pp. 1896–1909, 2005.

[9] S. Lombardi, N. K, Y. Makihara, and Y. Yagi, "Two-Point Gait: Decoupling Gait from Body Shape," in *ICCV*, 2013, pp. 1041–1048.

[10] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color Invariants for Person Re-Identification," *IEEE TPAMI*, vol. 35, no. 7, pp. 1622–1634, 2013.

[11] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person Re-identification by Iterative Re-weighted Sparse Ranking," *IEEE TPAMI*, pp. 1–1, 2014.

[12] C. Liu, S. Gong, and C. C. Loy, "On-the-fly Feature Importance Mining for Person Re-Identification," *Pattern Recognition*, 2013.

[13] B. Ma, Y. Su, and F. Jurie, "BiCov: a novel image representation for person re-identification and face verification," *BMVC*, pp. 57.1–57.11, 2012.

[14] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *ECCV*, 2006, pp. 125–136.

[15] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views," *CVIU*, vol. 109, no. 2, pp. 146–162, 2008.

[16] C.-T. Chu, J.-N. Hwang, K.-M. Lan, and S.-Z. Wang, "Tracking across multiple cameras with overlapping views based on brightness and tangent transfer functions," in *ICDSC*, no. 1, 2011, pp. 1–6.

[17] A. Datta, L. M. Brown, R. Feris, and S. Pankanti, "Appearance Modeling for Person Re-Identification using Weighted Brightness Transfer Functions," in *ICPR*, 2012.

[18] G. Zhang, Y. Wang, J. Kato, T. Marutani, and M. Kenji, "Local distance comparison for multiple-shot people re-identification," in *ACCV*, vol. 7726, 2013, pp. 677–690.

[19] S. Pedagadi, J. Orwell, and S. Velastin, "Local Fisher Discriminant Analysis for Pedestrian Re-identification," in *CVPR*, 2013, pp. 3318–3325.

[20] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[21] J. Fan, X. Shen, S. Member, Y. Wu, and S. Member, "What Are We Tracking : A Unified Approach of Tracking and Recognition," *IEEE TIP*, vol. 22, no. 2, pp. 549–560, 2013.

[22] N. Jiang, W. Liu, Y. Wu, and S. Member, "Learning Adaptive Metric for Robust Visual Tracking," *IEEE TIP*, vol. 20, no. 8, pp. 2288–2300, 2011.

[23] J. Yu, M. Wang, D. Tao, and S. Member, "Semisupervised Multiview Distance Metric Learning for Cartoon Synthesis," *IEEE TIP*, vol. 21, no. 11, pp. 4636–4648, 2012.

[24] A. C. Gallagher and C. Tsuhan, "Clothing cosegmentation for recognizing people," in *CVPR*, 2008, pp. 1–8.

[25] N. Gheissari, T. Sebastian, and R. Hartley, "Person Reidentification Using Spatiotemporal Appearance," in *CVPR*, 2006, pp. 1528–1535.

[26] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and Appearance Context Modeling," in *ICCV*, 2007, pp. 1–8.

[27] R. Garg, S. M. Seitz, D. Ramanan, and N. Snavely, "Where's Waldo: Matching people in images of crowds," in *CVPR*, 2011, pp. 1793–1800.

[28] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," in *AVSS*, 2011, pp. 179–184.

[29] S. Bak, E. Corvée, F. Brémond, and M. Thonnat, "Boosted human re-identification using Riemannian manifolds," *Image and Vision Computing*, vol. 30, no. 6-7, pp. 443–452, 2012.

[30] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *CVIU*, vol. 117, no. 2, pp. 130–144, 2013.

[31] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person Re-identification : What Features Are Important ?" in *ECCV*, 2012, pp. 391–401.

[32] Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao, "Collaborative Sparse Approximation for Multiple-Shot Across-Camera Person Re-identification," in *AVSS*, 2012, pp. 209–214.

[33] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human Re-identification by Matching Compositional Template with Cluster Sampling," in *ICCV*, no. 1, 2013, pp. 3152–3159.

[34] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-Based Person Re-Identification," in *AVSS*, 2013.

[35] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-Supervised Coupled Dictionary Learning for Person Re-identification," in *CVPR*, 2014.

[36] W. Li and X. Wang, "Locally Aligned Feature Transforms across Views," in *CVPR*, 2013, pp. 3594–3601.

[37] N. Martinel, C. Micheloni, and C. Piciarelli, "Learning pairwise feature dissimilarities for person re-identification," in *ICDSC*, 2013, pp. 1–6.

[38] N. Martinel, A. Das, C. Micheloni, and A. Roy-Chowdhury, "Re-Identification in the Function Space of Feature Warps," *IEEE TPAMI*, pp. 1–1, 2015.

[39] A. J. Ma, P. C. Yuen, and J. Li, "Domain Transfer Support Vector Ranking for Person Re-identification without Target Camera Label Information," in *ICCV*, 2013, pp. 3567–3574.

[40] M. Hirzer, P. M. Roth, K. Martin, and H. Bischof, "Relaxed Pairwise Learned Metric for Person Re-identification," in *ECCV*, 2012, pp. 780–793.

[41] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian Recognition with a Learned Metric," in *ACCV*, 2010, pp. 501–512.

[42] W. Li, R. Zhao, and X. Wang, "Human Reidentification with Transferred Metric Learning," in *ACCV*, 2012, pp. 31–44.

[43] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by Relative Distance Comparison." *IEEE TPAMI*, vol. 35, no. 3, pp. 653–668, 2013.

[44] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012, pp. 2666–2672.

[45] B. Ma, Y. Su, and F. Jurie, "Covariance Descriptor based on Bio-inspired Features for Person Re-identification and Face Verification," *IMAVIS*, vol. 32, pp. 379–390, 2014.

[46] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person Re-Identification by Regularized Smoothing KISS Metric Learning," *IEEE TCSVT*, vol. 23, no. 10, pp. 1675–1685, 2013.

[47] L. Ma, X. Yang, S. Member, and D. Tao, "Person Re-Identification Over Camera Networks Using Multi-Task Distance Metric Learning," *IEEE TIP*, vol. 23, no. 8, pp. 3656–3670, 2014.

[48] R. Zhao, W. Ouyang, and X. Wang, "Person Re-identification by Salience Matching," in *ICCV*, 2013, pp. 2528–2535.

[49] ——, "Unsupervised Salience Learning for Person Re-identification," in *CVPR*, 2013, pp. 3586–3593.

[50] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting Sparse Salient Regions." *IEEE TPAMI*, vol. 34, no. 1, pp. 194–201, 2011.

[51] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection." *IEEE TPAMI*, vol. 34, no. 10, pp. 1915–26, 2012.

[52] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain." *IEEE TPAMI*, vol. 35, no. 4, pp. 996–1010, 2013.

[53] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *NIPS*, 2007, pp. 554–552.

[54] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.

[55] K. Zhang, L. Zhang, and M.-h. Yang, "Real-Time Compressive Tracking," in *ECCV*, 2012, pp. 866–879.

[56] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[57] L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality Reduction: A Comparative Review," *JMLR*, vol. 10, no. February, pp. 1–41, 2009.

[58] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition and tracking," in *PETS*, 2007.

[59] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPeS: 3D People Dataset for Surveillance and Forensics," in *International ACM Workshop on Multimedia access to 3D Human Objects*, 2011, pp. 59–64.

[60] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[61] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for coding length increments," in *NIPS*, no. 800, 2008, pp. 681–688.

[62] D. Tao, L. Jin, Y. Wang, and X. Li, "Person Reidentification by Minimum Classification Error-Based KISS Metric Learning," *IEEE Transactions on Cybernetics*, pp. 1–11, 2014.

[63] B. Ma, Y. Su, and F. Jurie, "Local Descriptors Encoded by Fisher Vectors for Person Re-identification," in *ECCV*, 2012, pp. 413–422.

[64] M. Hirzer, P. M. Roth, and H. Bischof, "Person Re-identification by Efficient Impostor-Based Metric Learning," in *AVSS*, 2012, pp. 203–208.

[65] T. Zhou, M. Qi, J. Jiang, X. Wang, S. Hao, and Y. Jin, "Person Re-identification based on nonlinear ranking with difference vectors," *Information Sciences*, 2014.

[66] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *ICIP*, 2013, pp. 3567–3571.

[67] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person Re-Identification by Support Vector Ranking," in *BMVC*, 2010, pp. 21.1–21.11.