Saliency Weighted Features for Person Re-Identification

Niki Martinel, Christian Micheloni and Gian Luca Foresti

Department of Mathematics and Computer Science
University of Udine - 33100, Udine, Italy

Abstract. In this work we propose a novel person re-identification approach. The solution, inspired by human gazing capabilities, wants to identify the salient regions of a given person. Such regions are used as a weighting tool in the image feature extraction process. Then, such novel representation is combined with a set of other visual features in a pairwise-based multiple metric learning framework. Finally, the learned metrics are fused to get the distance between image pairs and to re-identify a person. The proposed method is evaluated on three different benchmark datasets and compared with best state-of-the-art approaches to show its overall superior performance.

1 Introduction

The person re-identification problem, i.e. identifying an individual across non-overlapping camera views, is becoming one of the most interesting tasks in computer vision. Many different applications, such as situational awareness [40], wide area scene analysis [39], etc. would benefit from it. This is supported by the relevant number of works presented by the community [46, 7].

The person re-identification problem is challenging due to different issues like variations on viewpoints and illumination conditions as well as persons poses and appearances. The non-rigid shape of the human body, background clutter and occlusions contribute to make the task non-trivial.

The computer vision community is addressing person re-identification by proposing both biometric-based (e.g. [52, 17, 45, 32]) and appearance-based methods (e.g. [25, 49, 43, 31]). Biometric-based methods exploit characteristic features (e.g. gait, face, etc.) of the person that are invariant in time but require a precise setup of the sensors (e.g. lateral or frontal view with relatively high spatial resolution). On the other hand, features extracted by appearance-based methods carry information about the appearance (i.e. clothes, carried on objects, etc.) which invariance is limited in time. But, these methods do not require a specific sensors configuration and deployment thus they are more suitable in context of wide area camera networks.

Solutions in the literature for re-identification can be further organized in three categories: (i) *Discriminative signature based methods* (e.g. [17, 25, 29, 33]) exploit human-defined person models that are matched using distance measures like L^2, χ^2, etc., or a combination of those. (ii) *Features transformation based methods* compute linear [14] and nonlinear [45, 23, 9, 10] transformation functions that are used to project features between different camera-dependent spaces. (iii) *Metric learning based algorithms* (e.g. [32, 49, 43, 24]) learn non-Euclidean distances that are used for the classification phase.

Motivation: Our solution builds upon three main considerations about the limits of current approaches:

- Most of the existing works compute the signature either directly from the whole image or silhouette, or by fusing local features extracted from dense image patches. This way each point of the person has the same importance in the computation of the signature. As [51, 50] we believe that the importance of the points is not uniform.
- Assuming we can compute the importance of points, it is not guaranteed that the same point is captured by all different views.
- Feature transformation functions are both highly non-linear [23, 9, 10] and depending on the class of the features, i.e., every feature transformation is modeled by a different function.

Contribution: We propose an approach that introduces three main novelties:

- A new kernelized graph-based technique to compute the importance of the points on the person, i.e., the saliency.
- The computed saliency is used as a weight in the feature extraction process: the higher the "saliency" the higher the importance of the feature for the re-identification and vice versa. This is combined with a set of local features to reduce the dependence on the saliency.
- A pairwise multiple metric learning framework used to model each feature space separately rather than jointly.

2 Related Work

Discriminative signature based methods seek for highly distinctive representations to describe a person's appearance under varying conditions. In [13], a region-based segmented image was used to extract spatio-temporal local features from multiple consecutive frames. In [12], a 2D rigid part based color appearance model was used to localize and match individuals in 3D system computed by means of the structure-from-motion technique. In [8, 6, 30, 36, 37], multiple local features were used to compute discriminative signatures for each person using multiple images. In [48], frames were used to build a collaborative representation that best approximates the query frames. In [4], Mean Riemannian Covariance patches extracted from individuals were used in a boosting scheme. In [25], re-identification was performed by matching shape descriptors of color distributions projected in the log-chromaticity space. In [51], an adjacency constrained patch matching strategy based on an unsupervised salient feature learning framework was used to improve re-identification accuracy. In [1], similarity with a reference set of persons was used in a Regularized Canonical Correlation Analysis framework. In [33], Biologically Inspired Features and covariance descriptors were used to compute the similarity between person images.

These methods addressed the problem by using human-defined person representations that are distinctive under changing conditions between different cameras. However, the exploited visual features are not be invariant to every variation that may affect the images acquired by disjoint cameras.

Features transformation based methods have addressed the re-identification problem by finding the transformation functions that affect the visual features acquired by disjoint cameras. In [23], a learned subspace of the computed brightness transfer function (BTF) between the appearance features was used to match persons across camera pairs. An incremental learning framework to model linear color variations between cameras has been proposed in [14]. In [9], the BTF was used to compensate the color difference between camera views. Tangent transfer functions derived by the homography between two cameras were also exploited to compensate the perspective difference. Usually the modeled functions are used to transform the feature space of one camera to the feature space of the other one. The re-identification then is performed in the so transformed feature space. Only recently, a few methods [2, 27, 38, 35] had also considered the fact that the transformation is not unique and it depends on several factors.

Metric learning based algorithms lie in between the two above categories as they still rely on particular features but they also advantage of a training phase to learn non-Euclidean distances used to compute the match in a different feature space. In [20], a relaxation of the positivity constraint of the Mahalanobis metric was proposed. In [11], unfamiliar matches were given less importance in the optimization problem in a Large Margin Nearest Neighbor framework. Multiple metrics specific to different candidate sets were learned in a transfer learning set up in [28]. In [49], the re-identification problem was formulated as a local distance comparison problem introducing an energy-based loss function that measures the similarity between appearance instances. In [24], a metric just based on equivalence constraints was learned. In [43], regularized Local Fisher Discriminant Analysis was introduced to maximize the between-class separability and preserve multi-class modality. In [31], it has been shown that user feedback solicited on-the-fly during the deployment stage for re-ranking boosts the re-identification performance over metric learning methods.

Recently, human saliency has also been explored in [51, 50]. However, in [51, 50] a reference set has been used, and, as claimed by the authors, the reference set is robust as long as its distribution well reflects the test scenario. To avoid this, we consider only neighborhood of pixels to compute the saliency. This brings three main benefits: (*a*) no need to find the best reference set for each context; (*b*) performance are not dependent from the reference set; (*c*) lower computational costs.

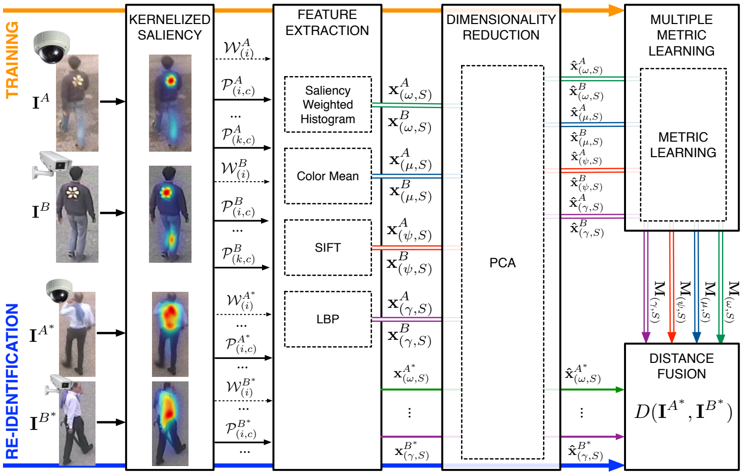

TRAINING
$\mathbf{I}^A$
$\mathbf{I}^B$
RE-IDENTIFICATION
$\mathbf{I}^{A^*}$
$\mathbf{I}^{B^*}$
KERNELIZED SALIENCY
$\mathcal{W}^A_{(i)}$
$\mathcal{P}^A_{(i,c)}$
...
$\mathcal{P}^A_{(k,c)}$
$\mathcal{W}^B_{(i)}$
$\mathcal{P}^B_{(i,c)}$
...
$\mathcal{P}^B_{(k,c)}$
$\mathcal{W}^{A^*}_{(i)}$
...
$\mathcal{P}^{A^*}_{(i,c)}$
...
$\mathcal{W}^{B^*}_{(i)}$
...
$\mathcal{P}^{B^*}_{(i,c)}$
...
FEATURE EXTRACTION
Saliency Weighted Histogram
Color Mean
SIFT
LBP
$\mathbf{x}^A_{(\omega,S)}$
$\mathbf{x}^B_{(\omega,S)}$
$\mathbf{x}^A_{(\mu,S)}$
$\mathbf{x}^B_{(\mu,S)}$
$\mathbf{x}^A_{(\psi,S)}$
$\mathbf{x}^B_{(\psi,S)}$
$\mathbf{x}^A_{(\gamma,S)}$
$\mathbf{x}^B_{(\gamma,S)}$
$\mathbf{x}^{A^*}_{(\omega,S)}$
⋮
$\mathbf{x}^{B^*}_{(\gamma,S)}$
DIMENSIONALITY REDUCTION
PCA
$\hat{\mathbf{x}}^A_{(\omega,S)}$
$\hat{\mathbf{x}}^B_{(\omega,S)}$
$\hat{\mathbf{x}}^A_{(\mu,S)}$
$\hat{\mathbf{x}}^B_{(\mu,S)}$
$\hat{\mathbf{x}}^A_{(\psi,S)}$
$\hat{\mathbf{x}}^B_{(\psi,S)}$
$\hat{\mathbf{x}}^A_{(\gamma,S)}$
$\hat{\mathbf{x}}^B_{(\gamma,S)}$
$\hat{\mathbf{x}}^{A^*}_{(\omega,S)}$
⋮
$\hat{\mathbf{x}}^{B^*}_{(\gamma,S)}$
MULTIPLE METRIC LEARNING
METRIC LEARNING
$\mathbf{M}_{(\gamma,S)}$
$\mathbf{M}_{(\psi,S)}$
$\mathbf{M}_{(\mu,S)}$
$\mathbf{M}_{(\omega,S)}$
DISTANCE FUSION
$D(\mathbf{I}^{A^*},\mathbf{I}^{B^*})$

Fig. 1. Proposed system architecture based on five main stages: kernelized saliency computation, feature extraction, dimensionality reduction, multiple metric learning and distance fusion.

3 The Approach

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

During training each image is given to the kernelized saliency detection module to compute the saliency of each pixel, i.e. a saliency map. Then, both the saliency map and the image are split into patches that are given to the feature extraction module. This module exploits the saliency as a weighting tool for the feature computation. Four different kind of features are extracted from all the patches and for each color component of the selected color spaces. Features of the same kind, extracted from the same color space, are concatenated and input to the dimensionality reduction module that applies Principal Component Analysis (PCA). Finally, the whole training set of PCA reduced features is given to the multiple metric learning module. This is in charge to learn a separate metric between two cameras for each feature type and color space.

During the re-identification phase, the same reduced representation computed for images coming from disjoint cameras, is given to the distance fusion module together with the learned metrics to compute the final dissimilarity.

3.1 Kernelized Saliency

Using our visual sensing field we are able to focus our limited perceptual and cognitive resources on the most pertinent subset of the available sensory data. We usually tell that a portion of the image is "salient" if it is "different" from its surroundings. However, being our goal to re-identify a target across disjoint cameras we have to deal with background clutter that may induce state-of-the-art saliency detection algorithms [22, 3, 47, 21, 44, 15, 26] to label “salient” a background region. But, we want only points on the person's silhouette to have high saliency.

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. Then, given a feature map $\mathbf{F} \in \mathbb{R}^{m \times n}$ we want to compute an activation map $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that features at coordinates (x, y), denoted as $\mathbf{F}_{(x,y)}$, which are “different” in their neighborhood and are close to the center of the image lead to high values of $\mathbf{A}_{(x,y)}$.

To achieve such objective the graph-based algorithm in [18] has been extended as follows. A fully-connected directed graph $G_{\mathbf{A}}$ is built by connecting every node (x, y) of $\mathbf{F}$ with all the other ones. The weight of each directed edge from node (x, y) to (p, q) is computed as

$$w((x,y),(p,q))_{\mathbf{A}} = \left| \log \left(\frac{\mathbf{F}_{(x,y)}}{\mathbf{F}_{(p,q)}} \right) \right| K_{\mathbf{F}} \left([x,y]^T, [p,q]^T \right) \tag{1}$$

where we have taken the standard definition of dissimilarity between the two feature values and weighted it by a kernel function $K_{\mathbf{F}}$ computed between the node locations. Once $G_{\mathbf{A}}$ has been computed, a Markov chain is defined over it in such a way that its equilibrium distribution accumulates mass at nodes that have high dissimilarities with their surroundings.

However, considering a single feature map $\mathbf{F}$ only limits the generality of the activation map. On the other hand, if multiple $\mathbf{A}$'s are computed we need to finally combine them into a single master activation map. This can be trivially solved using an additive combination. But, if each single activation map does not have similar mass at closer nodes, we may end up with a uniform and uninformative master map. Thus another fully connected graph $G_{\hat{\mathbf{A}}}$ is computed to concentrate the mass of each $\mathbf{A}$'s into nodes with high activation values. Let the weight between two nodes (x, y) and (p, q) be computed as

$$w((x,y),(p,q))_{\hat{\mathbf{A}}} = \mathbf{A}_{(p,q)} K_{\mathbf{A}} \left([x,y]^T, [p,q]^T \right) \tag{2}$$

where, as before, $K_{\mathbf{A}}$ is a kernel function. By defining a Markov chain over $G_{\hat{\mathbf{A}}}$ and normalizing the outbound edges to unity we can find the equilibrium distribution. As a result of this, we have that each concentrated activation map $\hat{\mathbf{A}}$ has most of the mass around nodes of $\mathbf{A}$ that have high activation values.

Now, let $\hat{\mathbf{A}}^j$ be the j-th activation map computed related to the $\mathbf{F}^j$ feature map, for $j = 1, 2, \ldots, J$. The final saliency map is defined as $\boldsymbol{\Omega} = \sum_{j=1}^{J} \boldsymbol{\alpha}_j \hat{\mathbf{A}}^j$ where $\boldsymbol{\alpha}$ is a vector of weights.

Three important characteristics have been achieved by introducing the kernel function in the above formulation: (i) the approach proposed in [18] has been generalized such that any kernel can be used to control the weight of the two graphs edges; (ii) the weight of the edge from node (x, y) to node (p, q) is kept proportional to their dissimilarity and to their closeness in the domain of $\mathbf{F}$; (iii) on average, nodes closer to the center have higher activation values than any particular point along the image boundaries. 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 somehow lies in the middle of the image.

3.2 Feature Extraction and Dimensionality Reduction

In the previous sections we introduced the idea that the points of a person's silhouette have different importance. However, it is not guaranteed that a very salient point acquired by a camera maintains the same property in a different camera (e.g. occluded point, change of pose, etc.). Nevertheless, if the same point is viewed by the two cameras and maintains the saliency properties then, it probably represents a good point in the re-identification process. To deal with both cases the saliency is used as a weight for a subset of features while the others are computed independently from the saliency.

As most state-of-the-algorithm methods the color, shape and texture features are considered in the proposed work. Before extracting such features, the RGB image $\mathbf{I}$ is projected into each color space $S \in \{HSV, Lab, YUV, rgs^1\}$. Then, the resulting image color channels $\mathbf{I}_{(c)}$, $c = 1, \dots, 12$, and the saliency map $\boldsymbol{\Omega}$ are divided into a set of k patches of equal size denoted $\mathcal{P}_{(i,c)}$ and $\mathcal{W}_{(i)}$ respectively, where $i = 1, \dots, k$ denotes the patch index.

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

$$\omega_{(i,c)}^{l,u} = \sum_{(x,y)\in\mathcal{P}_{(i,c)}} \begin{cases} \mathcal{W}_{(i,x,y)} & \text{if} \quad l < \mathcal{P}_{(i,c,x,y)} \leq u \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{W}_{(i,x,y)}$ and $\mathcal{P}_{(i,c,x,y)}$ are the saliency value and the pixel value at location (x, y) of patch i and color channel c. l and u are the lower and upper bin limits. (b) the color mean μ and (c) the 128-dimensional SIFT descriptor ψ. We also compute (d) the Local Binary Pattern (LBP) [42] 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 $\mathbf{x}_{(\omega,S)}$, $\mathbf{x}_{(\mu,S)}$, $\mathbf{x}_{(\psi,S)}$ and $\mathbf{x}_{(\gamma,gray)}$.

Notice that the saliency weight has been used in the extraction of histogram features only. This is due to the following reason. If we consider the task of re-identify a target across disjoint cameras, we are not guaranteed that a particular object/image region that has been assigned high saliency by one camera is visible by a different one. That is, we cannot rely on features extracted from those regions only to re-identify the target. Hence, we decided to weight only histogram features to balance this issue.

Due to the patch division the resulting feature vectors can be very high dimensional and each component may not have the same discriminative power. Principal Component Analysis is applied to each feature vector separately to get the vector of PCA coefficients $\hat{\mathbf{x}}_{(f,S)}$ where $f \in \{\omega, \mu, \psi, \gamma\}$ denotes the

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

feature type. To ease the notation, here and in the following, we use $S \in \{HSV, Lab, YUV, rgs, gray\}$. Notice that, while ω, μ and ψ are extracted from all the four selected color space, γ is computed only in the grayscale domain.

3.3 Multiple Metric Learning and Distance Fusion

For re-identification tasks, the input to metric learning algorithms is generally given by a vector representation of the image formed by concatenating multiple features (e.g. [49, 43, 24, 20]). Existing approaches have not considered that different kinds 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, in this work we exploit the algorithm proposed in [24] as it has no parameters that need to be optimized. We will introduce it briefly and show how the learned metrics can be fused to get the final dissimilarity.

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. As suggested in [24], by assuming that the feature space of pairwise differences is governed by a normal distribution (with zero mean) we can write the ration test as

$$\delta_{(f,S)}^{(A,B)} = \log\left(\frac{\mathcal{N}\left(\hat{\mathbf{x}}_{(f,S)}^{A} - \hat{\mathbf{x}}_{(f,S)}^{B}, \mathbf{0}, \mathbf{\Sigma}_{(A,B)=0}\right)}{\mathcal{N}\left(\hat{\mathbf{x}}_{(f,S)}^{A} - \hat{\mathbf{x}}_{(f,S)}^{B}, \mathbf{0}, \mathbf{\Sigma}_{(A,B)=1}\right)}\right) \tag{4}$$

where $\mathbf{\Sigma}_{(A,B)=1}$ and $\mathbf{\Sigma}_{(A,B)=0}$ are the sum of outer products computed for all the pairwise feature differences, $\hat{\mathbf{x}}_{(f,S)}^{A} - \hat{\mathbf{x}}_{(f,S)}^{B}$, that respectively belongs to the same person or to a different one.

Now, by taking the log of eq.(4) and discarding the constant terms that provide an offset, we can learn the Mahalanobis metric $\mathbf{M}_{(f,S)}$ by clipping the spectrum of $\hat{\mathbf{M}}_{(f,S)} = (\mathbf{\Sigma}_{(A,B)=1}^{-1} - \mathbf{\Sigma}_{(A,B)=0}^{-1})$ computed through eigenanalysis. Then, the Mahalanobis distance metric between the feature f extracted from the color space S of the images $\mathbf{I}^A$ and $\mathbf{I}^B$ is given by

$$d_{(f,S)}^2(\mathbf{I}^A, \mathbf{I}^B) = \left(\hat{\mathbf{x}}_{(f,S)}^{A} - \hat{\mathbf{x}}_{(f,S)}^{B}\right)^T \mathbf{M}_{(f,S)} \left(\hat{\mathbf{x}}_{(f,S)}^{A} - \hat{\mathbf{x}}_{(f,S)}^{B}\right). \tag{5}$$

The learned Mahalanobis distance metrics can then be fused to compute the final distance between person A and person B as

$$D(\mathbf{I}^A, \mathbf{I}^B) = \sum_f \sum_S \boldsymbol{\beta}_{(f,S)} d_{(f,s)}^2(\mathbf{I}^A, \mathbf{I}^B) \tag{6}$$

where $\boldsymbol{\beta}_{(f,S)}$ is a vector of distance weights.

3.4 Multiple Shot Extension

The re-identification community assumes that two sets of pedestrian images are available: the gallery set $\mathcal{G}$ (for which labels are known) and the probe set $\mathcal{T}$ (the set of pedestrians we want to re-identify). Let N be the number of images of each person in the two sets. Dependently on the value of N two matching philosophies are identified: i) single-shot ($N = 1$); ii) multiple-shot ($N > 1$). To extend our method to the multiple-shot scenario we take each feature f extracted from the color space S on N observations of a same person and pool them (mean operator). The reason for doing this is that the average of all observations is likely to be an estimate of the centroid for all samples and hence should be a valuable representation for each person.

4 Experimental Results

We evaluated our approach on three publicly available benchmark datasets: the VIPeR dataset [16], the 3DPeS dataset [5] and the CHUK02 dataset [28]. We chose these datasets as they provide many challenges faced in real world scenarios, i.e., viewpoint, pose and illumination changes, different backgrounds, image resolutions, occlusions, etc. More details about each dataset are discussed below.
Evaluation Criteria: We report 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 inside top k matches. On the other hand, the nAUC describes how well a method performs irrespectively of the dataset size. For each dataset, the evaluation procedure is repeated 10 times using independent random splits.
Implementation Details: To compute and fuse the saliency maps of an image we have taken the same settings in [18] and set both the kernel function $K_{\mathbf{F}}$ and $K_{\mathbf{A}}$ to be the standard Radial Basis Function with free parameter $\sigma = 1$. Each element of $\boldsymbol{\alpha}$ has been set to 1. We have sampled image patches of size 8×16 with a stride of 8×8 pixels to compute the weighted color histograms, each with 24 bins per channel. Similarly we have taken image patches of size 8×16 with a stride of 4×8 to compute the color mean and the LBP features. SIFT features have been extracted from 50% overlapping patches of size 16×16. Features have been reduced to $24, 46, 23, 21, 33, 33, 33, 33, 40, 26, 50, 34, 40$ dimensions. First 4 values are for histograms, second 4 are for the color means and the next 4 are for SIFT features extracted from the HSV (1st, 5th and 9th values), Lab (2nd, 6th and 10th values), YUV and rgs color spaces. Last value is for LBP features. Similarly, we set $\boldsymbol{\beta} = [0.06, 0.06, 0.01, 0.1, 0.1, 0.1, 0.1, 0.08, 0.02, 0.05, 0.1, 0.1, 0.1]$. All the parameters have been selected by 4-fold cross validation. Notice that, while these may have been separately estimated for each dataset, we have taken their average to provide a more general framework.

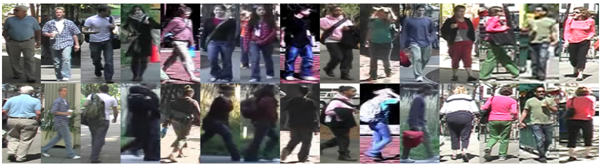


Fig. 2. 15 image pairs from the VIPeR dataset. The two rows show the different appearances of the same person viewed by two disjoint cameras.

4.1 VIPeR Dataset[2]

The VIPeR dataset [16] is one of the most challenging datasets for person re-identification due to the changes in illumination and pose, and the low spatial resolution of images. This dataset contains images of 632 persons viewed by two different cameras in an outdoor environment. Most of the image pairs show viewpoint changes larger than 90 degrees (see Fig. 2). To evaluate our method we followed the common protocol [43, 51, 1, 20] resizing all the images to 48×128.

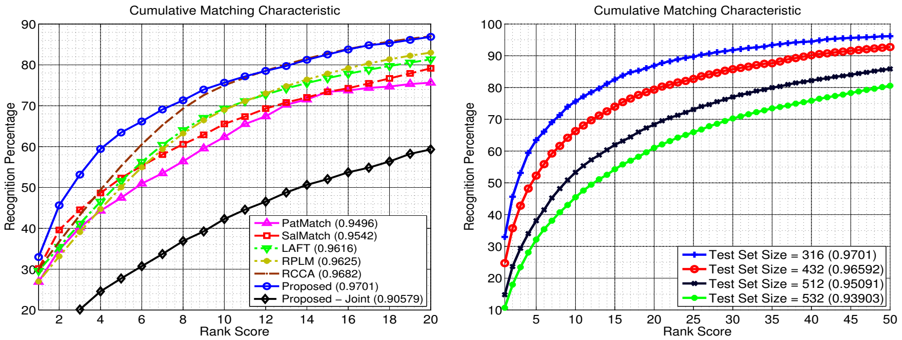
Table 1. Comparison with state-of-the-art methods on the VIPeR dataset. Best results for each rank are in boldface font. (*) Only results reported to 2 rounded digits are available. (**) The best run was reported, which cannot be directly compared to the other results.

Rank →	1	10	20	50	100	nAUC
Proposed	**32.97**	**75.63**	86.87	96.17	98.96	**0.9701**
SalMatch [50]	30.16	65.54	79.15	91.49	98.10	0.9542
RCCA(*) [1]	30	75	**87**	96	**99**	0.9682
LAFT [27]	29.60	69.30	81.34	**96.80**	-	-
RPLM(*) [20]	27	69	83	95	**99**	0.9625
PatMatch [50]	26.90	62.34	75.63	90.51	97.47	0.9496
eSDC.ocsvm [51]	26.74	62.37	76.36	-	-	-
eSDC.knn [51]	26.31	58.86	72.77	-	-	-
LF [43]	24.18	67.12	81.38	94.12	-	-
eLDFV [34]	22.34	60.04	71.00	88.92	**99**	0.9447
IBML(*) [19]	22	63	78	93	98	0.9516
CPS [8]	21.84	57.21	71.00	88.10	91.77	0.9360
eBiCOV [33]	20.66	56.18	68.00	84.90	88.66	0.9105
LMNN-R(**) [11]	20	68	80	93	**99**	0.9572

In Table 1 we compare our results to the state-of-the-art methods. Here we considered the scenario where half of the dataset is used for training and the remaining half is used for re-identification [3]. As shown, our method achieves the

[2] Available at `http://soe.ucsc.edu/~dgray/`

[3] Notice that some approaches are not using any training data as they're discriminative signature based methods (e.g. CPS [8], eBiCOV [33], etc.).



(a) (b)

Fig. 3. 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.

highest rank 1 score by reaching a recognition rate of 32.97% thus outperforming very recent and more complicated methods like SalMatch [50], RCCA [1] and LAFT [27]. We outperform the second runner up by more than 2.8%, which is a very important result on this dataset. From rank 1 to 10 we outperform all other methods by more than 6%. RCCA [1] is the only one, that, as our method, achieves a recognition percentage higher than 70% at rank 10. More interestingly, SalMatch [50], that has the second highest rank 1 score, achieves a recognition percentage of 65.54% for the same rank. For higher ranks (50 and 100) our method has very similar performance to the recent state-of-the-art methods. In general we achieve the best overall performance with an nAUC value of 0.9701.

In Fig. 3(a) we show the comparison of our method to the five top rank 1 approaches in terms of CMC curves. We also show that learning multiple metrics enables us to achieve better performance than learning a single metric for the joint feature space (black curve). To obtain the such result all the features have been concatenated, then PCA has been applied to reduce the feature space dimensions to 79 (best results found by cross-validation). At rank 10 a recognition rate of 75.63% is achieved by learning multiple metrics, while, concatenating the same features in a single vector and learning only one metric results in a recognition percentage of 42.16%.

In Fig. 3(b) we report the results of our method on the VIPeR dataset using different test set sizes. In Table 2, we also compare our method with RCCA [1], RPLM [20], PRDC [53], MCC [53], LAFT [27] and PCCA [41]. When the test set size contains 432 individuals we have the best performance on all the reported ranks. In particular we outperform the second runner up by more than 2.5% at rank 1 and by more than 7% when the considered rank is either 10 or 20. The same applies when the test set contains 512 individuals. We outperform all existing methods by more than 10% for rank 10 and 20. Finally, if we consider 532 persons as the test set, we achieve lower performance than RCCA [1] at rank

Table 2. Comparisons on the VIPeR dataset. Recognition rates per rank score as a function of the test set size.

Test Set Size	432			512				532		
Rank →	1	10	20	1	5	10	20	1	10	20
Proposed	**24.72**	**66.29**	**82.70**	**14.77**	**38.06**	**53.29**	**68.32**	10.67	45.46	**65.95**
RCCA [1]	22	59	75	-	-	-	-	**15**	**47**	60
RPLM [20]	20	56	71	-	-	-	-	11	38	52
PRDC [53]	13	44	60	9.12	24.19	34.40	48.55	9	34	49
MCC [53]	-	-	-	5.00	16.32	25.92	39.64	-	-	-
LAFT [27]	-	-	-	12.90	30.30	42.73	58.02	-	-	-
PCCA [41]	-	-	-	9.27	24.89	37.43	52.89	-	-	-

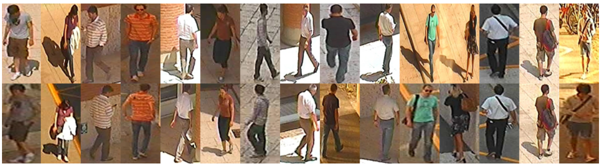


Fig. 4. 15 image pairs from the 3DPeS dataset. The two rows show the different appearances of the same person viewed by two disjoint cameras.

1. For higher ranks we achieve similar and superior performance than it and all other methods.

4.2 3DPeS Dataset[4]

The 3DPeS dataset [5] contains different sequences of 191 people taken from a multi-camera distributed surveillance system. There are 8 outdoor cameras and each one is presented with 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 shadowy areas. This results in a challenging dataset with strong variation of light conditions (see Fig. 4).

We compare our results to the ones reported in [43]. However, as in [43] 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 48 × 128 pixels. As this dataset comes with more than a single image per person per camera, we considered that all images have been used to compute the results in [43]. Then, as in [43] we randomly split the dataset into a training set and a test set containing 95 persons each.

In Fig. 5(a) we report the comparison of our method to three state-of-the-art approaches, namely LF [43], KISSME [24] and LMNN-R [11]. Our method achieves superior performance than all the others ones for ranks from 1 to 20.

[4] Available at `www.openvisor.org`

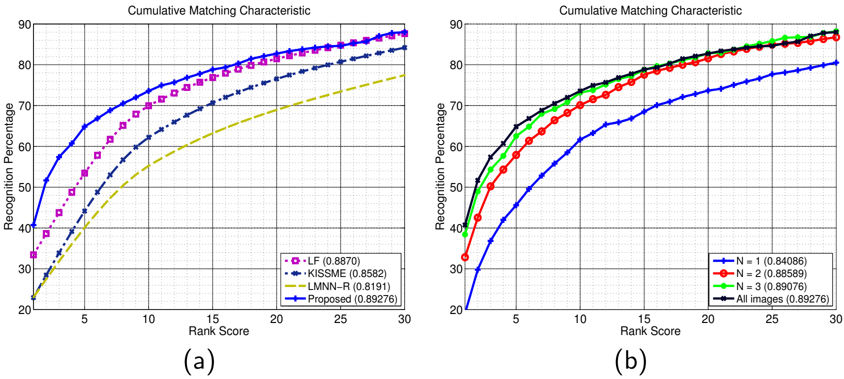


Fig. 5. Results on the 3DPeS dataset reported as averaged CMC curves. In (a) we compare our results to state-of-the-art methods: LF [43], KISSME [24] and LMNN-R [11]. In (b) we show the performance as a function of the number of shots used during both the training and the re-identification phase.

Table 3. Comparison of the proposed method on the 3DPeS dataset. Best results are in boldface font.

Rank →	1	10	25	50	nAUC
Proposed	**40.74**	**73.58**	84.64	94.90	**0.8928**
LF [43]	33.43	69.98	**84.80**	**95.07**	0.8870
KISSME [24]	22.94	62.21	80.74	93.21	0.8582
LMNN-R [11]	23.03	55.23	73.44	88.92	0.8191

In particular at rank 1 we achieve a correct recognition rate of 40.74% while, LF [43], KISSME [24] and LMNN-R [11] achieve a recognition rate of 33.43%, 22.94% and 23.03% respectively.

In Table 3 we report the comparison only for a subset of all the possible rank values. As shown, for lower ranks, our method outperforms the others. For higher ones, similar performance to LF [43] are achieved. Notice that the difference in performance for ranks 25 and 50 is less than 0.2%. As for the VIPeR dataset we reach the best overall performance with an nAUC value of 0.8928.

As the 3DPeS dataset comes with multiple images of the same pedestrian, in Fig. 5(b) we report the performance of our method as a function of N. In particular, as not all the persons come with an equal number of images, if the selected value of N is higher than the actual number of available images we take the maximum allowable number of images for that person. As shown in Fig. 5(b) when the single shot approach is considered ($N = 1$), our method achieves a recognition percentage of 19.05% at rank 1 and a recognition percentage of 73.68% when the considered rank is 20. At this rank our method outperforms LMNN-R [11] and meets the performance of KISSME [24] which have a recognition rate of 68.95% and 76.54% respectively. For these results LMNN-R [11] and KISSME [24] re-

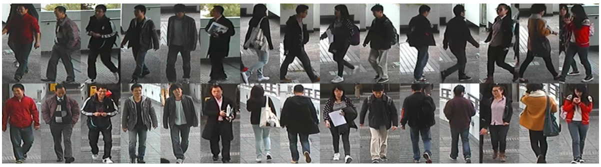


Fig. 6. 15 image pairs from the CUHK02 dataset. The two rows show the different appearances of the same person viewed by two disjoint cameras.

quire to use all the available images while we use a single one. Considering a multiple-shot modality, our method keeps constant the performance either using $N = 2$, $N = 3$ or all the available images. This is confirmed by the fact that the reported nAUC values change by less than 0.07 among all the three cases. However, by considering $N = 3$ (or all the available images) the rank 1 performance increases of about 6% with respect to the $N = 2$ scenario.

4.3 CUHK Campus Dataset[5]

The CUHK Campus dataset [28] 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 that occurs between camera pairs, this dataset is the one that has the highest number of persons collected by a single camera pair, thus it is the most representative for a real scenario. To evaluate our method and compare it to the state-of-the-art we follow the same protocol used in [28, 50]. Results are reported for camera pair P1 when $N = 2$ images per person are 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. 6). All the 3,884 images have been resized to 60 × 160. The dataset as been split into a training set containing 485 pedestrians and a test having images for the remaining 486.

In Fig. 7 we compare the results of our method to four state-of-the-art approaches, namely, SDALF [6], TML [28], PatMatch [50] and SalMatch [50]. At rank 1 our method performs better than all other ones by reaching a correct recognition rate of 31.05%, which improves the performance of SalMatch [50] by about 3%. Then, as the rank score increases our method outperforms all the other ones and, at rank 10, it achieves a recognition rate of 68.74%. As shown in Table 4, for the same rank SDALF [6], TML [28], PatMatch [50] and SalMatch [50] reach a recognition rate of 30.33%, 56.07%, 41.09% and 55.67%, respectively. Thus, for rank 10, we improve the state-of-the-art performance by more than 13%.

[5] Available at `http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html`

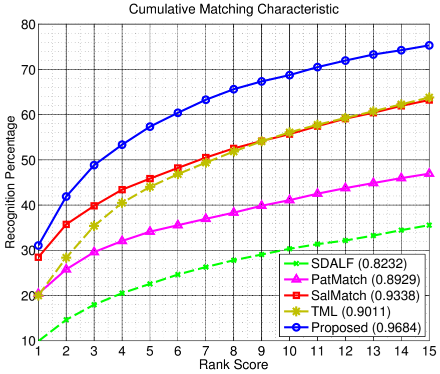


Fig. 7. Results on the CUHK02 Campus dataset (Camera P1) reported as averaged CMC curves. We show our superior performance to state-of-the-art approaches: SDALF [6], TML [28], PatMatch [50] and SalMatch [50].

Table 4. Comparison with state-of-the-art methods on the CUHK02 dataset (camera pair P1). Best results for each rank are in boldface font.

Rank →	1	5	10	20	50	100	200	nAUC
Proposed	**31.05**	**57.34**	**68.74**	**79.50**	**91.26**	**96.92**	**99.32**	**0.9684**
SalMatch [50]	28.45	45.85	55.67	67.95	83.52	92.10	98.10	0.9338
PatMatch [50]	20.39	34.12	41.09	51.56	68.42	82.00	93.23	0.8929
TML [28]	20.00	44.02	56.07	69.47	74.51	81.64	91.69	0.9011
SDALF [6]	9.90	22.57	30.33	41.03	55.99	67.39	84.12	0.8684

From the analysis of all the reported results, we can conclude that, in general, our method has superior performance than state-of-the-art approaches. This is supported by the fact that we achieve the best overall performance in terms of nAUC values for all the three considered datasets.

5 Conclusion

In this work we proposed to achieve the re-identification goal introducing a novel algorithm to identify the salient regions of a person. This is achieved by introducing a kernelized saliency that gives higher weights to the regions that are in the center of the image. Then, we used the computed saliency as a weight in a feature extraction process and combine it with other feature representations that do not consider it. The extracted features are used in a novel pairwise-based multiple metric learning framework. The learned metrics are finally fused to get the distance between image pairs and to re-identify a person. The novelty of the approach is also supported by the provided results. Our approach overall outperforms the best state-of-the-art solutions on the three most challenging benchmark datasets.

References

1. An, L., Kafai, M., Yang, S., Bhanu, B.: Reference-Based Person Re-Identification. In: Advanced Video and Signal-Based Surveillance (2013)
2. Avraham, T., Gurvich, I., Lindenbaum, M., Markovitch, S.: Learning Implicit Transfer for Person Re-identification. In: European Conference on Computer Vision, Workshops and Demonstrations. Lecture Notes in Computer Science, vol. 7583, pp. 381–390. Florence, Italy (2012)
3. Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): meaningful attention using stochastic image modeling. IEEE transactions on pattern analysis and machine intelligence 32(4), 693–708 (Apr 2010)
4. Bak, S., Corvée, E., Brémond, F., Thonnat, M.: Boosted human re-identification using Riemannian manifolds. Image and Vision Computing 30(6-7), 443–452 (Jun 2012)
5. Baltieri, D., Vezzani, R., Cucchiara, R.: 3DPeS: 3D People Dataset for Surveillance and Forensics. In: International ACM Workshop on Multimedia access to 3D Human Objects. pp. 59–64 (2011)
6. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding 117(2), 130–144 (Nov 2013)
7. Bedagkar-Gala, A., Shah, S.K.: A Survey of Approaches and Trends in Person Re-identification. Image and Vision Computing (Feb 2014)
8. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom Pictorial Structures for Re-identification. In: Procedings of the British Machine Vision Conference. pp. 68.1–68.11. British Machine Vision Association (2011)
9. Chu, C.T., Hwang, J.N., Lan, K.M., Wang, S.Z.: Tracking across multiple cameras with overlapping views based on brightness and tangent transfer functions. In: International Conference on Distributed Smart Cameras. pp. 1–6. No. 1, Ieee (Aug 2011)
10. Datta, A., Brown, L.M., Feris, R., Pankanti, S.: Appearance Modeling for Person Re-Identification using Weighted Brightness Transfer Functions. In: International Conference on Pattern Recognition. No. Icpr (2012)
11. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian Recognition with a Learned Metric. In: Asian conference on Computer vision. pp. 501–512 (2010)
12. Garg, R., Seitz, S.M., Ramanan, D., Snavely, N.: Where's Waldo: Matching people in images of crowds. In: International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1793–1800. IEEE (Jun 2011)
13. Gheissari, N., Sebastian, T., Hartley, R.: Person Reidentification Using Spatiotemporal Appearance. In: International Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1528–1535. IEEE (2006)
14. Gilbert, A., Bowden, R.: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: European Conference Computer Vision. pp. 125–136. Graz, Austria (2006)
15. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE transactions on pattern analysis and machine intelligence 34(10), 1915–26 (Oct 2012)
16. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recongnition, reacquisition and tracking. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). Rio De Janeiro, Brazil (Oct 2007)

17. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE transactions on pattern analysis and machine intelligence 28(2), 316–22 (Feb 2006)
18. Harel, J., Koch, C., Perona, P.: Graph-Based Visual Saliency. In: Advances in Neural Information Processing Systems. pp. 554–552. Vancouver (2007)
19. Hirzer, M., Roth, P.M., Bischof, H.: Person Re-identification by Efficient Impostor-Based Metric Learning. In: Advanced Video and Signal-Based Surveillance. pp. 203–208 (2012)
20. Hirzer, M., Roth, P.M., Martin, K., Bischof, H.: Relaxed Pairwise Learned Metric for Person Re-identification. In: European Conference Computer Vision. Lecture Notes in Computer Science, vol. 7577, pp. 780–793 (2012)
21. Hou, X., Harel, J., Koch, C.: Image Signature: Highlighting Sparse Salient Regions. IEEE transactions on pattern analysis and machine intelligence 34(1), 194–201 (Jul 2011)
22. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews. Neuroscience 2(3), 194–203 (Mar 2001)
23. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding 109(2), 146–162 (Feb 2008)
24. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: International Conference on Computer Vision and Pattern Recognition. pp. 2288–2295. No. Ldml (2012)
25. Kviatkovsky, I., Adam, A., Rivlin, E.: Color Invariants for Person Re-Identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(7), 1622–1634 (2013)
26. Li, J., Levine, M.D., An, X., Xu, X., He, H.: Visual saliency based on scale-space analysis in the frequency domain. IEEE transactions on pattern analysis and machine intelligence 35(4), 996–1010 (Apr 2013)
27. Li, W., Wang, X.: Locally Aligned Feature Transforms across Views. In: International Conference on Computer Vision and Pattern Recognition. pp. 3594–3601. IEEE (Jun 2013)
28. Li, W., Zhao, R., Wang, X.: Human Reidentification with Transferred Metric Learning. In: Asian Conference on Computer Vision. pp. 31–44 (2012)
29. Liu, C., Gong, S., Loy, C.C.: On-the-fly Feature Importance Mining for Person Re-Identification. Pattern Recognition (Nov 2013)
30. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person Re-identification : What Features Are Important ? In: European Conference on Computer Vision, Workshops and Demonstrations. pp. 391–401. Springer Berlin Heidelberg, Florence, Italy (2012)
31. Liu, C., Loy, C.C., Gong, S., Wang, G.: POP: Person Re-Identification Post-Rank Optimisation. In: International Conference on Computer Vision (2013)
32. Lombardi, S., K, N., Makihara, Y., Yagi, Y.: Two-Point Gait: Decoupling Gait from Body Shape. In: International Conference on Computer Vision. pp. 1041–1048 (2013)
33. Ma, B., Su, Y., Jurie, F.: BiCov: a novel image representation for person re-identification and face verification. British Machine Vision Conference pp. 57.1–57.11 (2012)
34. Ma, B., Su, Y., Jurie, F.: Local Descriptors Encoded by Fisher Vectors for Person Re-identification. In: European Conference on Computer Vision, Workshops and Demonstrations. pp. 413–422. Florence, Italy (2012)
35. Martinel, N., Micheloni, C.: Person re-identification by modelling principal component analysis coefficients of image dissimilarities. Electronics Letters 50(14), 1000–1001 (Jul 2014)

36. Martinel, N., Foresti, G.L.: Multi-signature based person re-identification. Electronics Letters 48(13), 765 (2012)
37. Martinel, N., Micheloni, C.: Re-identify people in wide area camera network. In: International Conference on Computer Vision and Pattern Recognition Workshops. pp. 31–36. IEEE, Providence, RI (Jun 2012)
38. Martinel, N., Micheloni, C., Piciarelli, C.: Learning pairwise feature dissimilarities for person re-identification. In: International Conference on Distributed Smart Cameras. pp. 1–6. IEEE, Palm Springs, CA (Oct 2013)
39. Martinel, N., Micheloni, C., Piciarelli, C., Foresti, G.L.: Camera Selection for Adaptive Human-Computer Interface. IEEE Transactions on Systems, Man, and Cybernetics: Systems 44(5), 653–664 (May 2014)
40. Micheloni, C., Remagnino, P., Eng, H.L., Geng, J.: Intelligent Monitoring of Complex Environments. IEEE Intelligent Systems 25(3), 12–14 (May 2010)
41. Mignon, A., Jurie, F.: PCCA: A new approach for distance learning from sparse pairwise constraints. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2666–2672. IEEE (Jun 2012)
42. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (Jul 2002)
43. Pedagadi, S., Orwell, J., Velastin, S.: Local Fisher Discriminant Analysis for Pedestrian Re-identification. In: International Conference on Computer Vision and Pattern Recognition. pp. 3318–3325 (2013)
44. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: International Conference on Computer Vision and Pattern Recognition. pp. 733–740. Ieee (Jun 2012)
45. Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R.: Matching shape sequences in video with applications in human movement analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12), 1896–1909 (Dec 2005)
46. Vezzani, R., Baltieri, D., Cucchiara, R.: People Re-identification in Surveillance and Forensics: a Survey. ACM Computing Surveys 46(2) (2014)
47. Wang, W., Wang, Y., Huang, Q., Gao, W.: Measuring visual saliency by Site Entropy Rate. In: International Conference on Computer Vision and Pattern Recognition. pp. 2368–2375. No. 2, IEEE (Jun 2010)
48. Wu, Y., Minoh, M., Mukunoki, M., Li, W., Lao, S.: Collaborative Sparse Approximation for Multiple-Shot Across-Camera Person Re-identification. In: Advanced Video and Signal-Based Surveillance. pp. 209–214. Ieee (Sep 2012)
49. Zhang, G., Wang, Y., Kato, J., Marutani, T., Kenji, M.: Local distance comparison for multiple-shot people re-identification. In: Asian conference on Computer Vision. Lecture Notes in Computer Science, vol. 7726, pp. 677–690 (2013)
50. Zhao, R., Ouyang, W., Wang, X.: Person Re-identification by Salience Matching. In: International Conference on Computer Vision. pp. 2528–2535. IEEE (Dec 2013)
51. Zhao, R., Ouyang, W., Wang, X.: Unsupervised Salience Learning for Person Re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3586–3593. IEEE (Jun 2013)
52. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Surveys 35(4), 399–458 (2003)
53. Zheng, W.S., Gong, S., Xiang, T.: Re-identification by Relative Distance Comparison. IEEE transactions on pattern analysis and machine intelligence 35(3), 653–668 (Jun 2013)