Classification of Local Eigen-Dissimilarities for Person Re-Identification

Niki Martinel*, *Student Member, IEEE,* and Christian Micheloni, *Member, IEEE,*

***Abstract*—The task of re-identifying a person that moves across cameras fields-of-view is a challenge to the community known as the person re-identification problem. State-of-the art approaches are either based on direct modeling and matching of the human appearance or on machine learning-based techniques. In this work we introduce a novel approach that studies densely localized image dissimilarities in a low dimensional space and uses those to re-identify between persons in a supervised classification framework. To achieve the goal: i) we compute the localized image dissimilarity between a pair of images; ii) we learn the lower dimensional space of such localized image dissimilarities, known as the "local eigen-dissimilarities" (LEDs) space; iii) we train a binary classifier to discriminate between LEDs computed for a positive pair (images are for a same person) from the ones computed for a negative pair (images are for different persons). We show the competitive performance of our approach on two publicly available benchmark datasets.**

***Index Terms*—Person re-identification, pairwise appearance modeling, eigen-representation**

I. Introduction

Countless work have been proposed by the community to address the problem of tracking pedestrian (e.g. [1], [2], [3], [4], [5]) within a single camera field-of-view (FoV), i.e. intra-camera tracking. However, in a very large camera network, not all the areas can be covered by the deployed sensors [6], [7]. As a result of this, we have to deal with the person re-identification problem, formally defined as the task of assigning the same label to a person that moves across camera FoVs, i.e. inter-camera tracking with no temporal constraints. Such problem is very attractive as, for video surveillance applications, knowing whether a person is present in the monitored area at a precise time instant is of paramount importance. This is supported by the relevant recent works in [8], [9]

To address the challenging issues of the person re-identification problem, the community has devoted effort following three main approaches. *Discriminative signature based methods* have been the most widely used ones. In [10], an unsupervised approach to learn the most discriminating features was proposed. In [11], color distributions were investigated to identify the color-invariant intra-distribution structure. In [12], Biologically Inspired Features were used to compute the similarity between images. In [13], re-identification was performed by measuring similarity with a reference dataset in a Regularized Canonical Correlation Analysis (CCA) subspace.

Other approaches have addressed the re-identification problem by modeling the *transformation of features* between pairs of cameras. In [14], the Brightness Transfer Function (BTF) computed between the appearance features was used to match persons across camera pairs. An incremental learning framework to model linear color variations between cameras was proposed in [15]. In [16], the BTF was used to compensate the color difference between camera views. In [17], dissimilarities between multiple features extracted from image pairs were used to train a binary random forest classifier. In [18], image spaces of two camera views were split into different configurations according to the similarity of cross-view transforms. Then, the transformation of features was separately learned for each of the split spaces.

Finally, only recently, *metric learning based algorithms* have been introduced in the field of person re-identification. In [19], the re-identification problem was formulated as a local distance comparison problem. In [20], a metric based on equivalence constraints was learned. Such a metric was extended in [21] by adding a smooth regularizer. In [22], a relaxation of the positivity constraint of the Mahalanobis metric was introduced. Multiple metrics specific to different candidate sets were learned in a transfer learning set up in [23]. Similarly, in [24], different metrics were learned for different feature types. In [25], Principal Component Analysis (PCA) and Local Fisher Discriminant Analysis were applied in a metric learning framework.

Motivation and contribution: All of such methods try to capture the most discriminative characteristics of each person by means of local or global color, shape and texture features. While this has been shown to be an effective approach [10], extracting such features is a complex and computationally expensive process that makes them not suitable for real-time scenarios. We also believe that single image pixels carry more discriminative information than such complex features.

Comparing pixels at exactly the same location suffers due to the variability in a person's pose and location. To deal with this issue, we group a set of neighboring pixels to have a coarse representation. Then, we build upon the idea that there exists a multi-modal transformation of the difference between such localized pixel groups. However, not all the localized groups may be useful to capture such multi-modal transformation. Therefore, the core contribution of this work is to capture the multi-modal transformation of the difference between groups of localized pixels that lie on a linear subspace that best captures the intrinsic dimensionality of those. Towards this objective, we compute the local dissimilarity between

The authors are with the Department of Mathematics and Computer Science, University of Udine, Udine, Italy, 33100, Italy. E-mail: {niki.martinel, christian.micheloni}@uniud.it

Manuscript received June 28, 2013; revised August 1, 2013.

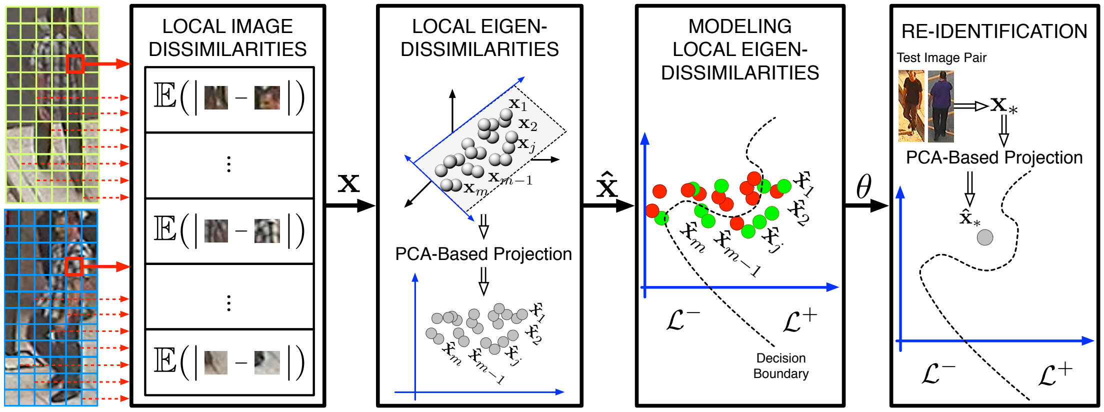


Fig. 1. Proposed system overview. Local image dissimilarities are computed for each of the patches into which the given images are split. Then PCA projection is applied to get the LEDs and the magnitude of them along the new basis is used to learn the parameter of the decision boundary that separates the set of positive and negative LEDs. (*Best viewed in color.*)

images of the same person (positive pair) as well as the ones between images of different persons (negative pair) viewed in two cameras. Then, we show that the set composed of all the local dissimilarities lies in a linear subspace that can be learned using a Principal Component Analysis (PCA)-based algorithm. Finally, we use a supervised classification framework to discriminate between the positive and negative pairs in the linear subspace. The approach is evaluated using two benchmark datasets showing competitive results compared to the state-of-the-art.

II. Methodology

An overview of our approach is shown in Fig. 1. The key idea is to address the re-identification by discriminating between pair of images of the same person from those computed for images of different persons in the space of local eigen-dissimilarities (LED). Towards this objective, we first divide the image into a set of overlapping patches. Then, for each of these we compute the image dissimilarities and learn the space formed by all LEDs. Finally, we take the magnitudes of the LEDs along the new basis and use these as features in an offline binary classification framework.

A. Local Image Dissimilarities

To tackle the re-identification challenges, all state-of-the-art methods for person re-identification have explored different image representations by using appearance features [10]. However, in this process, part of the information in the original image may be lost. To address this, we consider the LED between two images as a feature for re-identification.

Let $\mathbf{I}_A$ and $\mathbf{I}_B$ be the images of two persons, A and B, acquired by two disjoint cameras. For each image, we first take a set of dense patches $\{\mathbf{P}_i(\mathbf{I}) \in \mathbb{R}^{M \times N}\}_{i=1}^{n}$, where n is the total number of dense patches[1]. Then we compute their dissimilarities as

$$\mathbf{d}_i = |\mathbf{P}_i(\mathbf{I}_A) - \mathbf{P}_i(\mathbf{I}_B)| \tag{1}$$

where $\mathbf{d}_i \in \mathbb{R}^{MN}$ and $|\cdot|$ is the absolute value function.

If we separately consider the dissimilarity of each pixel in the two patches as in Eq.(1), even for patches with a relatively small spatial resolution, we end up with a very high dimensional feature space [2]. To avoid this, and to capture most of the discriminating power of the patch, we apply the expectation operator $\mathbb{E}$ to each $\mathbf{d}_i$ (see Fig. 2) to get the local dissimilarity (LD) vector $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ where $x_i = \mathbb{E}(\mathbf{d}_i)$ for all $i = 1, \cdots, n$.

B. Unsupervised Learning of Eigen-Dissimilarities

The expectation operator applied on each dense patch enables us to (i) capture and reduce the high-dimensionality of the image dissimilarity, and (ii) smoothen the noise out. However, as we assume that no silhouette extraction has been performed, we are considering patches that belong to the background to compute the LD. Thus, it is not necessary to use every component of the LD vector to discriminate between the set of positive and negative pairs.

Most of the existing algorithms based on image dissimilarities (e.g. [22], [21]) assume that their representations lie in a linear subspace. To model such subspace, they generally adopt the standard PCA technique. However, applying PCA on the LD and retaining the largest eigenvalues is intuitively wrong. Consider a common scenario in which background clutter and occlusions are present [1]. A non-matching pixel between an image pair (i.e., a pixel belonging to the person in one view and on the background in the other view) have large variations in the LD due to the changing background. Likewise, a matching element between the two images (i.e., an element corresponding to the same part of the person) tends introduce small variations in the LD.

On the basis of all such considerations we exploit PCA as follows. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m]$ be the matrix formed by m LDs. Assume that each component in X has zero mean. Then, by applying PCA to $\mathbf{X}$ we end up with the matrix $\hat{\mathbf{X}} \in \mathbb{R}^{(n-k)\times m}$, where k denotes the number of principal components (i.e. those with largest eigen-values) that

[1] In our current framework we use the same settings in [13], [22]. We sample $N = 8$ by $M = 16$ patches with 50% overlap in both directions.

[2] Let consider an image $\mathbf{I} \in \mathbb{R}^{128\times 64}$ and let divide it into patches with $M = 16$ and $N = 16$ pixels with 50% of overlapping between these. By computing the pixel-by-pixel difference for all the $n = 105$ patches we end up with a concatenated feature vector of 26880 dimensions.

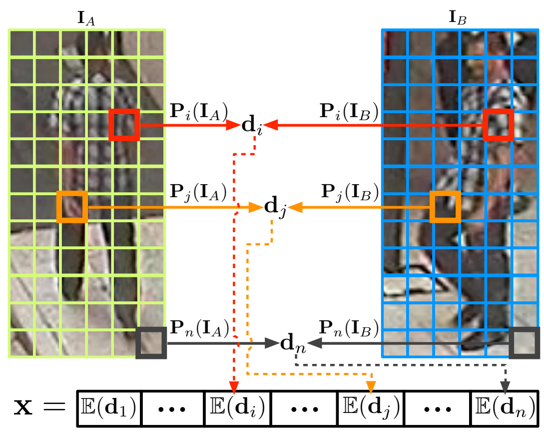


Fig. 2. Computation of the local image dissimilarities.

are rejected [3]. Each column vector $\hat{\mathbf{x}}_i$ in $\hat{\mathbf{X}}$ is now the vector of non-principal PCA coefficients, that is the LED vector.

C. Supervised Re-Identification

Let $\mathcal{L}$ be the LED space composed by $\mathcal{L}^+$, the set of all positive LEDs, i.e. the LEDs computed for pairs of images of the same person, and $\mathcal{L}^-$, the set of all negative LEDs, i.e. the LEDs computed for pairs of images of different persons. We use a Support Vector Machine classifier with a Radial Basis Function kernel to learn a mapping from the LED space $\mathcal{L}$ to the label space, $y = \{-1, +1\}$.

Let m be the number of LED training examples and $\hat{\mathbf{x}}_i$ for $i = 1, \cdots, m$, then, the objective of the SVM learning procedure in the dual form can be defined as

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{m} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{j,k} \boldsymbol{\alpha}_j \boldsymbol{\alpha}_k y_j y_k K(\hat{\mathbf{x}}_j, \hat{\mathbf{x}}_k, \sigma) \right\} \quad (2)$$

$$\text{s.t.} \quad 0 \leq \boldsymbol{\alpha}_i \leq C, i = 1, \cdots, m \quad \wedge \quad \sum_{i=1}^{m} \boldsymbol{\alpha}_i y_i = 0$$

where C is a regularization parameter, controlling the penalty for imperfect fit to training labels, and $K(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j, \sigma)$ is the standard Radial Basis Function kernel with free parameter σ.

Let $\boldsymbol{\theta}$ characterize the set of SVM parameters that maximizes the objective function in Eq.(2) and let $\hat{\mathbf{x}}_*$ be the test LED computed using images of person A and person B. Then, to tell whether A and B are the same person, we compute the posterior probability $P(y = 1 | \hat{\mathbf{x}}_*; \boldsymbol{\theta})$ using the commonly adopted Platt scaling.

The community poses the re-identification problem by assuming that two sets of persons images are available: the gallery set $\mathcal{G}$ (for which labels are known) and the probe set $\mathcal{P}$ (the set of persons we want to re-identify) [26]. This gives rise to two matching philosophies: i) single-shot, when only one image of a person is present in each set; ii) multiple-shot, when both $\mathcal{G}$ and $\mathcal{P}$ contain multiple images of a person.

To support both the single-shot and multiple-shot mechanisms [8] we do the following. Let n_A be the number of images of person A in one camera and n_B be the number of images of person B in a different camera. To get a final probability of A and B being the same person we take the average of all the $n_A \times n_B$ probabilities.

[3]To reject the noisy LD components, in our approach we select k such that 95% of the variance is rejected.

Fig. 3. A set of 15 randomly taken images pairs from the CAVIAR4REID dataset.

III. Experimental Results

We evaluate the performance our method using two publicly available benchmark datasets: the CAVIAR4REID [4] and the 3DPeS [5] dataset. Each of the two dataset introduces particular challenges that make them a suitable choice for evaluating person re-identification methods. A description of each of those is given in the following.

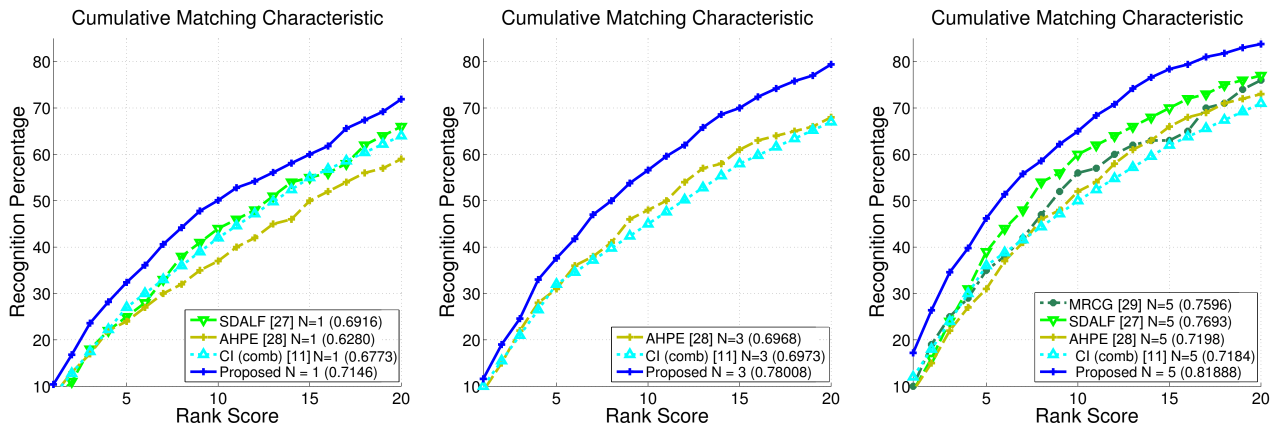
We report our results as Cumulative Matching Characteristic (CMC) curve and the normalized Area Under Curve (nAUC) for both a single-shot and multiple-shot strategy. We define the number of images for each person as $N = n_A = n_B$. In our current framework, the LDs are computed between *rgs* color[6] images and the SVM parameters have been selected performing a grid search by 4-fold cross validation.

A. CAVIAR4REID Dataset

The CAVIAR4REID dataset has 1220 images of 72 persons out of which 50 are acquired by two non-overlapping cameras. The images are of different sizes, varying from 39×17 to 144×72 with illumination and pose changes (see Fig. 3). For evaluation on this dataset, we followed the procedure described in [11], i.e., we only consider the 50 persons that are viewed by the two cameras. As other methods, we assume that all the 50 persons appear both in the training as well as in the test set, but we take disjoint image samples for the two phases. The whole procedure is repeated 10 times and the average performance are reported.

In Fig. 4 we compare the results of our method with those achieved by SDALF[27], AHPE [28], CI (comb) [11] and MRCG [29] considering both the single-shot ($N = 1$) and multiple-shot ($N \in \{3, 5\}$) scenarios. For the single shot scenario we achieve better performance to the methods. This is supported by the fact that our method has the best nAUC value (0.7146). Considering 2 more images to build the gallery set and the probe set significantly improves the results. In fact, for the multiple-shot scenario with $N = 3$, we outperform all other methods by reaching a correct recognition of about 58% at rank 10, while the same result is achieved at rank 14 and 15 for AHPE and CI (comb) respectively. For $N = 5$ we achieve a rank 1 correct recognition percentage of 17%. SDALF, AHPE, CI (comb) and MRCG achieve a correct recognition percentage of 9%, 9% 12%, and 10% for the same

[4]Available at http://www.lorisbazzani.info/code-datasets/caviar4reid/
[5]Available at http://www.openvisor.org/3dpes.asp
[6]$r = R/(R + G + B), g = G/(R + G + B), s = (R + G + B)/3$



(a) (b) (c)

Fig. 4. Comparison of the proposed algorithm to state-of-the-art methods for person re-identification on the CAVIAR4REID dataset. In 4(a) results are shown for the a single-shot while in 4(b) and 4(c) results are reported for two multiple-shot scenarios (with $N = 3$ and $N = 5$ respectively).

Fig. 5. A set of 15 randomly taken images pairs from the 3DPeS dataset.

TABLE I
COMPARISONS ON THE 3DPES DATASET. BEST RESULTS ARE IN BOLDFACE FONT.

Rank Score	1	10	25	50	nAUC
Proposed (N=1)	27.84	66.74	82.42	94.21	0.8735
Proposed (N=3)	**35.35**	**73.91**	**86.20**	**95.78**	**0.896**4
LF [25]	33.43	69.98	84.80	95.07	0.8870
KISSME [20]	22.94	62.21	80.74	93.21	0.8582
LMNN-R [30]	23.03	55.23	73.44	88.92	0.8191

rank 1 respectively. In this case we are the only one having an nAUC value higher than 0.8.

B. 3DPeS Dataset

The 3DPeS dataset has been proposed in [31]. It contains different sequences of 191 people taken from a multi-camera distributed surveillance system. The pedestrians were detected multiple times with different viewpoints, at different time instants, in clear light and in shadow areas (see Fig. 5).

We evaluate our method using the protocol defined in [25]. We partitioned the dataset into a training set and a test set containing 95 persons each. However, in [25] it was not clear how many images per person were used to compute the reported results (i.e. the value of N). In light of this, we used both the single-shot and the multiple-shot modalities to report our results and compare those to LF [25], KISSME [20] and LMNN-R [30]. The results are reported in terms of CMC and nAUC values averaged over 10 different trials.

As shown in Table I, if we consider the single-shot scenario we achieve better performance than KISSME and LMNN-R, but not of LF. On the other hand, if we consider the multiple-shot scenario ($N = 3$) we perform better than all such methods by reaching a rank 1 recognition rate of 35.35%. The same performance behavior is reflected on all the other 4 provided ranks. In particular, our method is the only one that has a recognition higher than 70% at rank 10. This shows that, while being simple, our method is able to handle hard re-identification cases and performing better than much more complex methods like LF.

TABLE II
COMPUTATIONAL PERFORMANCE ANALYSIS - TIMES REQUIRED TO EXTRACT THE PROPOSED FEATURES FROM A SINGLE IMAGE.

Algorithm	Feature Extraction/Modeling Time [ms]
Proposed	122.37
RPML [22]	920.03
KISSME [20]	885.60

C. Computational Performance

To show that our method can be used in real-time scenarios, we have conducted studies on the feature extraction process[7]. In Table II we report the computational time required to extract the LED features from a given image. These are compared to the feature extraction times required by [22], [20]. Results are averaged over the 1012 images in the 3DPeS dataset. Notice that, while LED are computed for an image pair, the other algorithms extract features from each single image separately. Hence, they require twice the reported times to model an image pair. Results demonstrate that our method significantly decreases the computational performance with respect to other state-of-the-art methods.

IV. CONCLUSION

In this work, we tackled the person re-identification problem by introducing a novel method that studies the nature of the localized error between images of the same person and images of two different persons. We have shown that the localized errors lie in the linear subspace of “local eigen-dissimilarities”. We have used the representation in the new subspace as a feature in a binary classification framework. Experimental results carried out on two benchmark datasets have shown the benefit of such representation. Overall, better performance than state-of-the-art has been achieved at lower computational effort.

[7]Experiments have been carried out using a non-optimized MATLAB code running on an Intel i7, Windows x64, 16GB RAM machine.

REFERENCES

[1] Y. Yi and H. Xu, "Hierarchical Data Association Framework with Occlusion Handling for Multiple Targets Tracking," *IEEE Signal Processing Letters*, vol. 21, no. 3, pp. 288–291, Mar. 2014.

[2] D. Wang, H. Lu, and C. Bo, "Online Visual Tracking via Two View Sparse Representation," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1031–1034, Sep. 2014.

[3] Y. Xu, L. Qin, G. Li, and Q. Huang, "Online Discriminative Structured Output SVM Learning for Multi-Target Tracking," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 190–194, Feb. 2014.

[4] C. Wang, H. Liu, and Y. Gao, "Scene-Adaptive Hierarchical Data Association for Multiple Objects Tracking," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 697–701, Jun. 2014.

[5] N. Martinel, C. Micheloni, C. Piciarelli, and G. L. Foresti, "Camera Selection for Adaptive Human-Computer Interface," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 653–664, May 2014.

[6] C. Micheloni, P. Remagnino, H.-L. Eng, and J. Geng, "Intelligent Monitoring of Complex Environments," *IEEE Intelligent Systems*, vol. 25, no. 3, pp. 12–14, May 2010.

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

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

[9] A. Bedagkar-Gala and S. K. Shah, "A Survey of Approaches and Trends in Person Re-identification," *Image and Vision Computing*, Feb. 2014.

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

[11] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color Invariants for Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

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

[13] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-Based Person Re-Identification," in *Advanced Video and Signal-Based Surveillance*, 2013.

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

[15] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *European Conference Computer Vision*, Graz, Austria, 2006, pp. 125–136.

[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 *International Conference on Distributed Smart Cameras*, no. 1. Ieee, Aug. 2011, pp. 1–6.

[17] N. Martinel, C. Micheloni, and C. Piciarelli, "Learning pairwise feature dissimilarities for person re-identification," in *International Conference on Distributed Smart Cameras*. Palm Springs, CA: IEEE, Oct. 2013, pp. 1–6.

[18] W. Li and X. Wang, "Locally Aligned Feature Transforms across Views," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2013, pp. 3594–3601.

[19] G. Zhang, Y. Wang, J. Kato, T. Marutani, and M. Kenji, "Local distance comparison for multiple-shot people re-identification," in *Asian conference on Computer Vision*, ser. Lecture Notes in Computer Science, vol. 7726, 2013, pp. 677–690.

[20] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *International Conference on Computer Vision and Pattern Recognition*, no. Ldml, 2012, pp. 2288–2295.

[21] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person Re-Identification by Regularized Smoothing KISS Metric Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.

[22] M. Hirzer, P. M. Roth, K. Martin, and H. Bischof, "Relaxed Pairwise Learned Metric for Person Re-identification," in *European Conference Computer Vision*, ser. Lecture Notes in Computer Science, vol. 7577, 2012, pp. 780–793.

[23] W. Li, R. Zhao, and X. Wang, "Human Reidentification with Transferred Metric Learning," in *Asian Conference on Computer Vision*, 2012, pp. 31–44.

[24] N. Martinel, C. Micheloni, and G. L. Foresti, "Saliency Weighted Features for Person Re-Identification," in *ECCV Workshops and Demonstrations*, no. i, 2014, pp. 1–17.

[25] S. Pedagadi, J. Orwell, and S. Velastin, "Local Fisher Discriminant Analysis for Pedestrian Re-identification," in *International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[26] N. Martinel, C. Micheloni, and C. Piciarelli, "Distributed Signature Fusion for Person Re-Identification," in *International conference on Distributed Smart Cameras*, Hong Kong, Hong Kong, 2012, pp. 1–6.

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

[28] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, Nov. 2011.

[29] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," in *International Conference on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, 2011, pp. 179–184.

[30] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian Recognition with a Learned Metric," in *Asian conference on Computer vision*, 2010, pp. 501–512.

[31] 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.