Re-Identification in the Function Space of Feature Warps

Niki Martinel, *Student Member, IEEE*, Abir Das, *Student Member, IEEE*, Christian Micheloni, *Member, IEEE*, and Amit K. Roy-Chowdhury, *Senior Member, IEEE*

Abstract—Person re-identification in a non-overlapping multicamera scenario is an open challenge in computer vision because of the large changes in appearances caused by variations in viewing angle, lighting, background clutter, and occlusion over multiple cameras. As a result of these variations, features describing the same person get transformed between cameras. To model the transformation of features, the feature space is nonlinearly warped to get the "warp functions". The warp functions between two instances of the same target form the set of feasible warp functions while those between instances of different targets form the set of infeasible warp functions. In this work, we build upon the observation that feature transformations between cameras lie in a nonlinear function space of all possible feature transformations. The space consisting of all the feasible and infeasible warp functions is the warp function space (WFS). We propose to learn a discriminating surface separating these two sets of warp functions in the WFS and to re-identify persons by classifying a test warp function as feasible or infeasible. Towards this objective, a Random Forest (RF) classifier is employed which effectively chooses the warp function components according to their importance in separating the feasible and the infeasible warp functions in the WFS. Extensive experiments on five datasets are carried out to show the superior performance of the proposed approach over state-of-the-art person re-identification methods. We show that our approach outperforms all other methods when large illumination variations are considered. At the same time it has been shown that our method reaches the best average performance over multiple combinations of the datasets, thus, showing that our method is not designed only to address a specific challenge posed by a particular dataset.

Index Terms—Feature transformation, Person re-identification, warp function space

1 Introduction

With the advancement of imaging sensor technology, surveillance systems have seen remarkable increase in various application areas ranging from home surveillance to small business and large retail applications, from facility access and environment monitoring, to open border surveillance. Even though the sensing devices are becoming cheaper, monitoring a wide area by deploying a large number of cameras is still not feasible due to the amount of human supervision, privacy concerns, and maintenance costs involved. As a result, only a small part of the whole area is covered by a number of cameras with non-overlapping fields-of-view (FoVs). The non-overlapping camera FoVs leave *"blind gaps"* which are critical in the sense that no information can be obtained from these areas. This raises the need for automated methods able to extract, and access high-level semantic information carried by the extremely high volume of recorded video data. As a result of losing a person when he/she leaves a camera FoV, it is extremely challenging to re-associate the same person at a different location and time among multiple persons. This inter-camera person association problem is known as the person re-identification problem.

In spite of a surge of effort put in by the research community in recent years, re-identification has remained quite an open issue due to a number of hard challenges. Firstly, footages are recorded in an uncontrolled environment by cameras with large FoVs, generating low resolution images of the targets. 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 perform unsatisfactorily. As a result, visual appearance features are, still, the first choice in re-identification problems. As a target's appearance often undergoes large variations across non-overlapping camera views due to significant changes in viewing angle, lighting, background clutter, and occlusion, the appearance features for the target can be very different from camera to camera. This is especially true in case of person re-identification due to the non-rigid shape of the human body. An example of such a scenario is shown in Fig. 1. Three frames of the same person acquired by three non-overlapping cameras are presented together with the color histogram (hue, saturation and value) features. As shown, such features are significantly different for the same target viewed in different cameras making the re-identification problem very challenging.

The computer vision community has tried to address the re-identification problem by designing discriminative signatures for each target or by finding a non-euclidean metric

- *N. Martinel and C. Micheloni are with the Department of Mathematics and Computer Science, Univeristy of Udine, Udine 33100, Italy. E-mail: {niki.martinel, christian.micheloni}@uniud.it.*
- *A. Das and A.K. Roy-Chowdhury are with the Department of Electrical Engineering, University of California Riverside, Riverside, CA 92521. E-mail: abir.das@email.ucr.edu, amitrc@ee.ucr.edu.*

Manuscript received 20 Nov. 2013; revised 28 Aug. 2014; accepted 13 Nov. 2014. Date of publication 3 Dec. 2014; date of current version 6 July 2015. Recommended for acceptance by S. Sarkar.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2377748

Fig. 1. Three images of the same person in three non-overlapping cameras from the RAiD dataset [1]. Below each image, HSV features are shown as three different histograms. Brown denotes the hue, green denotes saturation and the sky-blue denotes the value histograms respectively. The inconsistency of the histogram shape does not allow them to be used as unique features for re-identification.

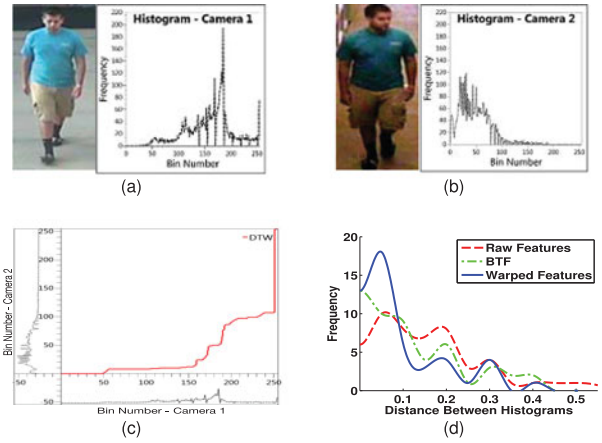

Fig. 2. Using the principle of DTW to capture the transformation of features as a person goes from a brightly illuminated space to a dark place. (a) and (b) show the images of a person along with its value histogram plots at a brightly illuminated and a dark place respectively. (c) shows the warp function which maps the bin numbers of the color histogram in (a) to the bin numbers of the color histogram in (b). The initial flatness and latter steepness of the warp function captures the transformation of features resulting from the change in illumination. (d) shows the distribution of the Bhattacharyya distances between the transformed and actual grayscale histograms using BTF [2] (in green) and warp functions (in blue) computed for all the 50 persons in the CAVIAR4REID dataset. Concentration of more persons with smaller distances using warp function can be readily seen. The distribution of the distances computed between the raw value histograms is also shown for comparison (in red).

which minimizes distance between the features of the same target across cameras. These methods rely on the fact that the individual signatures vary a little from camera to camera. Such methods, while efficient and effective to re-identify persons viewed in different poses, result in a significant loss of performance when strong illumination and color changes occur between different cameras. As a result of these changes, features describing the same person get transformed between cameras. Thus an important aspect of the problem is to understand how features get transformed across cameras. Fig. 2 shows an example where a person goes from a brightly illuminated space (Fig. 2a) to a dark place (Fig. 2b). This large change of illumination is also depicted by the shift of the distribution of pixels from the higher end values towards the lower end values in the corresponding grayscale histograms (shown alongside the two images). This change in the shape of the distribution can be captured by studying the histogram warp. We use the principles of dynamic time warping (DTW) for this purpose. DTW [3], [4] is a dynamic programming algorithm that optimizes the alignment of two time series by non-linearly warping the series so that the sum of the point-to-point distances is minimized. Time sequences are functions of time while color histograms are functions of the bin numbers. So the same principle can be used to study the warping of the bin number axis causing the change in the shape of the distributions. Fig. 2c shows such a *warp function* which captures the feature transformation by mapping the bin numbers of the color histogram in Fig. 2a (shown as the horizontal axis) to the bin numbers of the color histogram in Fig. 2b (shown as the vertical axis). The initial flatness and latter steepness of the warp function characterize the shift of the concentration of the pixels from the higher to the lower end of the color histogram. Fig. 2d shows a comparative performance of the use of warp function and an widely used feature transformation method, the brightness transfer function (BTF) [2] on capturing the feature transformation. Value histogram features of images from one camera in CAVIER4RAID [5] dataset was transformed to features from another camera using warp functions and BTFs. Bhattacharyya distances between the transformed feature and the original feature in the second camera are computed for both the feature transformation methods. As shown by Fig. 2d, the distribution of the number of people for which the distance is smaller is more when warp function is used than when BTF is used to transform the feature from one camera to other.

The existing studies exploiting feature transformation, have tried to learn linear [6] and nonlinear transformation functions [2], [7] between appearance features among pairs of cameras. These approaches, however, use the learned transformation function to project the features from one camera to the feature space of the other camera. In a re-identification scenario this may not always be feasible since the mapping may not be unique and it may vary from frame to frame depending on a large number of camera parameters (e.g. illumination, scene geometry, exposure time, focal length, and aperture size). In this work, we build upon a detailed understanding of the transformation of features captured by warp functions computed based on the principles of DTW. Considering two non-overlapping cameras, a pair of images of the same target is denoted as a feasible pair, while a pair of images between two different targets is denoted as an infeasible pair. The corresponding warp functions describing the transformation of features are denoted as *feasible* (positive) and *infeasible* (negative) warp functions respectively. The set of infeasible warp functions vary widely as in this set the warps are computed for image pairs consisting of different persons. Even within the set of feasible warp functions, the transformations are not unique when computed for different feasible pairs. For each of the two sets, the feature transformations may not be well represented by a single warp function in presence of such variabilities. So, we propose to model the function space capturing all the feasible and infeasible warp functions between pairs of cameras, termed as the feature *warp*

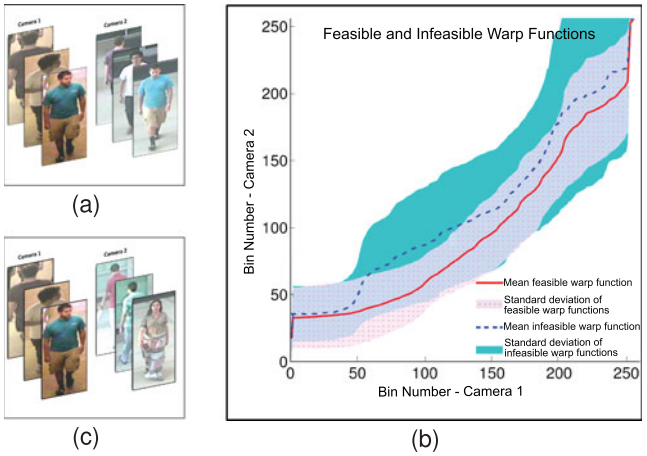

Fig. 3. Feasible and infeasible warp functions in the WFS. (a) and (c) show example images of the feasible and infeasible pairs respectively taken from an outdoor and an indoor camera of the RAiD [1] dataset. (b) shows the mean of the feasible (in bold line) and infeasible warp functions (in dashed line) between the grayscale histograms of the torso of the feasible and infeasible pairs. 100 randomly chosen examples of feasible and infeasible warp functions are averaged to get the mean warp functions. The shaded areas show the corresponding spread of the variances (as $\pm$ standard deviation value). This figure shows that feasible and infeasible warp functions for this simple feature (grayscale histogram) can be discriminative and can be used for re-identification.

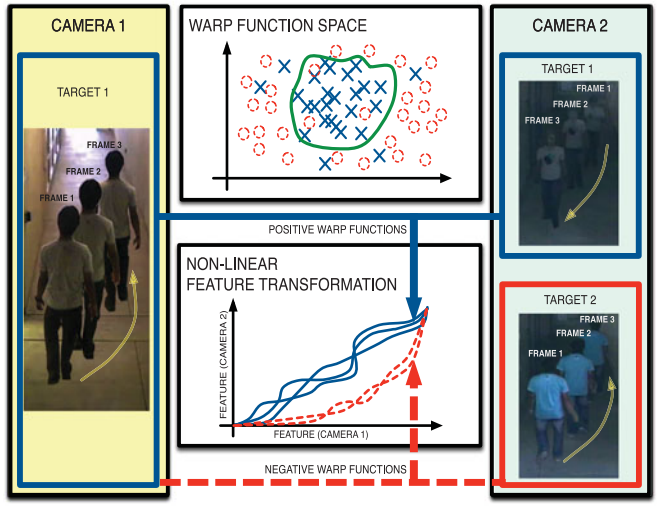

Fig. 4. Re-identification by discriminating in the warp function space. The warp functions computed between features extracted from images of the same target (i.e. positive warp functions) are shown in solid blue. The warp functions computed between features extracted from different targets (i.e. negative warp functions) are shown in dashed red. A nonlinear decision surface (shown in green) is learned to separate the two regions.

function space (WFS). The WFS not only allows us to model feasible transformation between pairs of instances of the same target, but also to separate them from the infeasible transformations between instances of different targets. This enables us to address the re-identification problem as a binary classification problem by discriminating in the WFS.

Fig. 3 shows a visual proof of the discriminating power of the feasible and infeasible warp functions. For convenience of visualization, we resorted to a low dimensional WFS by computing the warp functions between the grayscale histograms of the images. Figs. 3a and 3c respectively show some examples of feasible and infeasible image pairs from the RAiD [1] dataset corresponding to camera 1 and 3. Since, in general, people wear different colored clothes for torso and legs the warp functions for the two bodyparts are computed separately. For visualization convenience we show the warp functions for torso only. Fig. 3b shows the mean feasible (in bold line) and infeasible warp functions (in dashed line) between the grayscale histograms of 100 randomly chosen feasible and infeasible pairs of images respectively. The shaded areas show the corresponding spread of the variances (as $\pm$ standard deviation value). This shows that both the mean warp function and the spread of variance are different for feasible and infeasible warp functions even for this simple feature (grayscale histogram). The proposed work explores this discriminating power of the feasible and infeasible warp functions in the WFS for person re-identification. Since, most of the benchmark datasets include changes of scale and viewpoint in addition to illumination, it may not always be possible to discriminate well enough using such a simple feature representation. So, we computed the warp functions between other dense color and texture features in the actual experimentations to deal with these challenges. Discrimination between the two classes of warp functions are further enhanced in a classification framework which finds a complex discriminating surface in a higher dimensional WFS consisting of the warp functions computed between these features. Details of the feature extraction and computation of warp functions can be obtained in Section 4.

To summarize, the contributions of the proposed approach to the problem of person re-identification are the followings. To capture the feature transformation we propose to compute a nonlinear mapping (warp function) that minimizes a cost defined as the mismatch between histogram features. A WFS composed of the collection of feasible and infeasible warp functions is built. We also propose to learn a discriminating surface between the sets of feasible and infeasible warps in the WFS using a random forest (RF) of decision trees. The re-identification problem is addressed by mapping a test warp function onto the WFS and classifying it as belonging to either the set of feasible or infeasible warp functions (see Fig. 4).

We compare the performance of our approach to state-of-the-art person re-identification methods using five publicly available benchmark datasets. The datasets are chosen with a particular focus on large illumination variation between cameras. Since we learn the space of feature transformations, our results significantly outperform others when applied to datasets with large appearance variations between the cameras, such as RAiD [1] and WARD [8]. Also, we demonstrate that our method is not tuned to any specific dataset. Our average performance on different combinations of multiple datasets is higher than other state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 gives a description of the state-of-the-art approaches in person re-identification. An overview of the proposed approach is given in Section 3. The details about the re-identification approach, as feature extraction, warping and WFS are described in Section 4. Experimental results and comparisons with state-of-the-art methods are shown in Section 5. Finally, conclusions are drawn in Section 6.

2 Previous Works in Person Re-Identification

In the last few years the problem of re-identifying persons across multiple non-overlapping cameras has received increasing attention. The community has commonly adopted three different kind of approaches: i) discriminative signatures based methods, ii) metric learning based methods, iii) transformation learning based methods. A multidimensional taxonomy and categorization of the person re-identification algorithms can be obtained in the review paper [9]. In the rest of the section we do a thorough review of the existing re-identification works.

Discriminative signature based methods [5], [8], [10], [11] used multiple standard features e.g., color, shape, texture etc. or specilly learned features like biologically inspired features (BIF) [12], covariance descriptors [13], shape descriptors of color distributions in log-chromaticity space [14] etc. to compute discriminative signatures for each person using multiple images. Some recent methods have shown that the adoption of salient feature learning approaches [15] and representing the query images based on reference datasets [16], [17] can be used to boost the re-identification performance.

According to [18], in a metric learning framework a set of training data is used to learn a non-euclidean metric which minimizes the distance between features of pairs of true matches while maximizing the same between pairs of wrong matches. Works trying to improve the metric learning performance by excluding well separable examples and solving an eigenvalue problem [19], by giving less importance to unfamiliar matches in a large margin nearest neighbor framework [18], by learning multiple metrics specific to different candidate sets in a transfer learning set up [20] or by exploiting sparse pairwise similarity/dissimilarity constraints [21] have shown remarkable re-identification performance. Metric learning based person re-identification has also been formulated as a local distance comparison problem on energy-based loss functions [22], [23] or Local Fisher Discriminant Analysis [24]. To reduce the computational costs, a relaxation of the positivity constraint of the Mahalanobis metric has been proposed [25]. A detailed description of metric learning approaches is beyond the scope of the work. The interested readers are directed to two survey papers [26], [27] on this subject. A similar approach to metric learning is dissimilarity measure learning [28] which has been used successfully in person re-identification [29], [30]. These methods create a set of dissimilarity descriptors based on a set of visual prototypes obtained by unsupervised clustering. Person re-identification is, then, formulated as a supervised classification problem with the learned dissimilarity descriptors as features. A basic difference of the metric learning or dissimilarity measure based methods with our approach is that these methods do not take into account the transformation of features which is especially useful when there is a significant but consistent change of appearance of the individuals between cameras. Also the methods based on person specific signature, dissimilarity measure and metric learning have to either rely on the assumption that all the persons are seen during the training phase or carefully choose threshold value separating the new persons from the matches with existing persons. Since we are exploiting transformation of features between cameras and it is independent of the specific persons, the proposed method is more general in this sense.

In one of the early works [7] studying the transformation of features, a BTF between appearance features was computed by finding the optimal path in the feature correlation matrix. Later, a learned subspace of the computed BTFs [2] and a incremental learning framework modeling linear color variations [6] between cameras were used to match the targets. Both [6] and [2] learned space-time probabilities of moving targets between cameras and used them as cues for association. However, transition time information may be unreliable if camera FoVs are significantly non-overlapping. Efforts of improving the BTF resulted in a BTF modeling the effects of illumination changes over time [31], a sparse color information preserving Cumulative BTF [32], or a Weighted BTF designed to assign unequal weights to observations based on how close they are to test observations [33]. In [34] the re-identification problem was posed as a classification problem in the feature space formed of concatenated features of persons viewed in two different cameras.

In this work we focus specifically on the issue of how features are transformed between views and learn a model of these transformation functions. We pose the re-identification problem as computing these nonlinear warp functions between features and learning a function space which models the feasible and the infeasible warp functions.

3 Overview of Proposed Approach

The overall scheme of the proposed person re-identification process is shown in Fig. 5. Given the frames from two cameras we learn a discriminative model in the WFS to get the probability of a sample feature warp function coming from the same person or not.

Towards this objective, we first extract features from the person images. The feature extraction module performs the following tasks: a) splitting the image of the detected persons into four main body parts, and b) extracting dense color and texture features from the detected body parts.

For each extracted feature, vector valued warp functions are computed by the warp function space module. All the warp functions (corresponding to different features) are concatenated to form a high dimensional warp function for each image pair. The warp function between the same target in different cameras is denoted as a feasible or positive warp function while the warp function between two different targets is denoted as an infeasible or a negative warp function. The set of all feasible and infeasible warp functions forms the WFS. The dimensionality of the WFS is reduced using principal component analysis (PCA) [35].

Given the WFS, a decision surface discriminating the two sets of warp functions is learned using a random forest [36] of bagged decision trees. Every component of the warp functions may not be discriminating enough between the two classes of transformations (feasible/infeasible). The decision trees select the subset of warp function components according to their importance and maximize the discrimination between the feasible and infeasible warp functions in the WFS.

For classification, features are extracted from test image pairs and input to the WFS module to compute the warp

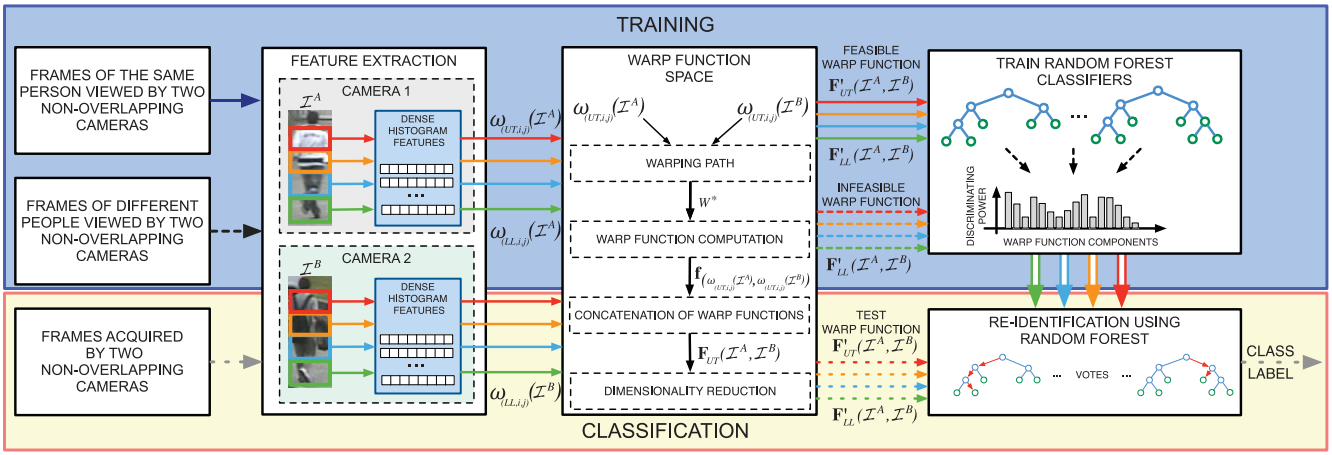

Fig. 5. System overview. The feature extraction module takes raw video frames and extracts dense color and texture features from each of the four detected body parts. These are input to the warp function space module that computes the warp function between each of them and reduces the dimensionality of the warp function space. A random forest classifier is trained to discriminate between the feasible and the infeasible warp functions in the WFS. The trained classifier is used to classify the test warp functions.

functions. Finally, the RF classifies the test warp functions in the WFS as feasible or infeasible.

4 Methodology

In this section we describe the different modules of the proposed approach in details.

4.1 Feature Extraction

State-of-the-art methods for person re-identification have successfully explored different appearance features [11]. While existing feature transformation based methods are designed for color features, our framework can be used to study the nature of the transformation of any feature which, in turn, can be used for re-identification. In this work we focus not only on color features but also on popular texture features.

Before computing these features, we identify the salient regions like head $\mathcal{I}_H$, torso $\mathcal{I}_T$ and legs $\mathcal{I}_L$ from the given image $\mathcal{I}$ as proposed in [10]. In our approach we only consider $\mathcal{I}_T$ and $\mathcal{I}_L$ since the head region $\mathcal{I}_H$ often consists of a few and less informative pixels. We additionally divide both $\mathcal{I}_T$ and $\mathcal{I}_L$ into two horizontal sub-regions based on the intuition that people can wear shorts or long pants and short or long sleeves tops. The four different regions are resized to fixed height and width to extract fixed size dense features from all of them. We denote these resized regions as $\hat{\mathcal{I}}_\phi$ where $\phi \in \{UT, LT, UL, LL\}$ denotes the upper-torso, lower-torso, upper-legs and lower-legs regions respectively. The resized regions are further divided into non overlapping patches $\mathcal{P}_{(\phi,1)}, \mathcal{P}_{(\phi,2)}, \ldots, \mathcal{P}_{(\phi,n_\phi)}$ of size $R \times R$ each, where n_ϕ denotes the number of patches corresponding to the body part ϕ. Then, for all the patches $\mathcal{P}_{(\phi,i)}$, $i = 1, \ldots, n_\phi$ we extract the following features.

Color. State-of-the-art person re-identification methods use color features relying on the assumption that persons do not change their clothes as they move between camera FoVs. According to that, and following the considerations in [11], we exploit the HSV, CIELab, RGB and YCbCr color spaces to extract the dense histogram features. For image $\mathcal{I}$, bodypart ϕ and patch i we extract the histogram $\omega_{(\phi,i,c)}(\mathcal{I}) \in \mathbb{R}^{b_c}$, where b_c is the number of bins of the feature histogram for color component $c \in \{H, S, V, L, a^*, b^*, R, G, B, Y, Cb, Cr\}$.

Texture. Similar to color features, we extract dense texture features to capture the appearance of a person. We use LBP texture feature which is computationally efficient and is robust to both gray-scale variations [37] and rotation [38]. The extracted LBP texture histogram is denoted as $\omega_{(\phi,i,LBP)}(\mathcal{I}) \in \mathbb{R}^{b_{LBP}}$, where b_{LBP} is the number of bins used to quantize the resulting LBP histogram. We also use Gabor [39], Schmid [40] and Leung-Malik (LM) [41] filter banks to extract texture features. After convolving the ith patch with each filter of the filter banks we compute the modulus of the response and quantize it in histograms of b_G, b_{Schmid} and b_{LM} bins respectively for the above three filter banks. Denoting the set of individual filters in Gabor, Schmid and LM filter banks as G, S and LM, the set of color and texture features extracted from patch $\mathcal{P}_{(\phi,i)}$ is given by the set $\{\omega_{(\phi,i,j)}(\mathcal{I})\}$ where $j \in \{c \cup LBP \cup G \cup S \cup LM\}$. An example of the responses of such filter banks is shown in the supplementary, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2014.2377748. Fig. 6 shows an example image where dense features from the four bodyparts have been extracted as described above.

4.2 Warp Function Space

To capture the transformation of the extracted features between cameras, we use the principles of Dynamic Time Warping. DTW [42] has been widely used in many fields such as speech recognition [43], data mining [44], activity recognition [45], [46] etc. DTW finds patterns that govern change of shape from one time series to another. This dynamic programming based algorithm non-linearly warps the time axis of a time series so that it is optimally aligned to the other time series with minimum cost of alignment. The cost is, in general, the sum of the point to point-to-point distances of the two time series elements. Time sequences are functions of time while feature histograms are functions of the bin numbers. In our approach the bin number axis is warped to reduce the mismatch between feature values of two feature histograms from two cameras.

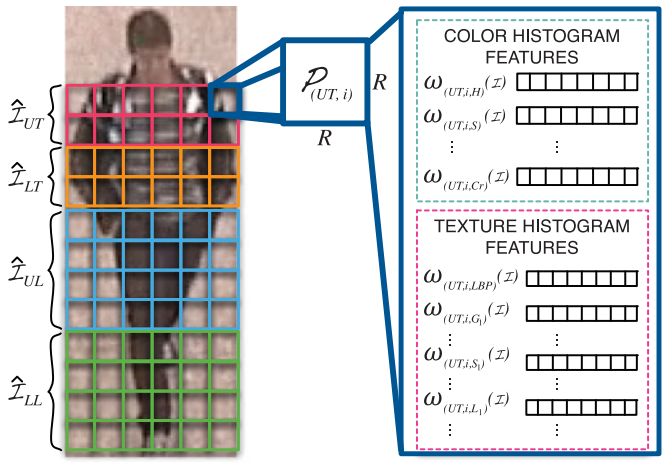

Fig. 6. Dense image features from the detected body parts. Dense color and texture histogram features are extracted from each of the four resized body parts.

Let $\mathbf{x}(1,\ldots,m) = \langle x(1),\ldots,x(m)\rangle$ and $\mathbf{y}(1,\ldots,m) = \langle y(1),\ldots,y(m)\rangle$ be two vector valued functions. Let f be a warp function from $\mathbf{x}$ to $\mathbf{y}$, that is

$$y(a) = x(f(a)), f(a) : [1,m] \rightarrow [1,m] \in \mathcal{F}, \tag{1}$$

where $\mathcal{F}$ is the space of all warp functions, the WFS.

To find the warp function, a cost matrix $C \in \mathbb{R}^{m\times m}$ is generated where the (a,b)th element (denoted as C_{ab}) of the matrix is given by the distance $\delta(x(a),y(b)), \forall a,b \in \{1,2,\ldots,m\}$. Though any suitable distance function can be used or learned using a metric learning procedure, in general, the magnitude of the difference and the euclidean distance between elements are adopted due to their simplicity [3]. The warp function is the path giving the lowest cumulative cost between fixed start point, the $(1,1)$th cell and fixed end point, the (m,m)th cell of C. Let $\mathbb{W} = \{W_1, W_2, \ldots\}$ be the set of all possible paths between these two fixed points where W_q denotes the qth path. W_q consists of tuples indicating the indices of the cells in C. Then the optimal warp path is given by,

$$W^* = \underset{W_q\in\mathbb{W}}{\operatorname{argmin}} \left(\sum_{(a,b)\in W_q} C_{ab} \right). \tag{2}$$

The optimization problem in (2) is solved in a dynamic programming framework under suitable monotonicity and continuity constraints [3], [4]. Finding the non-linear warp path W^* does not guarantee that the length of the warp path is same for all feature pairs $\mathbf{x}$ and $\mathbf{y}$. This is due to the fact that the mapping $f(a) : \{1,2,\ldots m\} \rightarrow \{1,2,\ldots m\}$, described by the tuples in W^* is, in general, many to many. To get a m length warp function we employ the following rule for all $(a,b) \in W^*$

$$f(a) = \begin{cases} \min(b) & \text{if } a \neq 1,\ m, \\ a & \text{otherwise.} \end{cases} \tag{3}$$

Gathering the $f(a)$'s for all $a = 1,2,\ldots,m$ in a vector $\mathbf{f}_{(\mathbf{x},\mathbf{y})}(1,\ldots,m) = \langle f(1),\ldots,f(m)\rangle$ we get the warp function that warps $\mathbf{x}$ to $\mathbf{y}$.

In our approach the warp function $\mathbf{f}$ is computed for each feature and for every dense patch (see Section 4.1). In other words, as shown in Fig. 7, $\mathbf{f}$ is computed for feature pairs

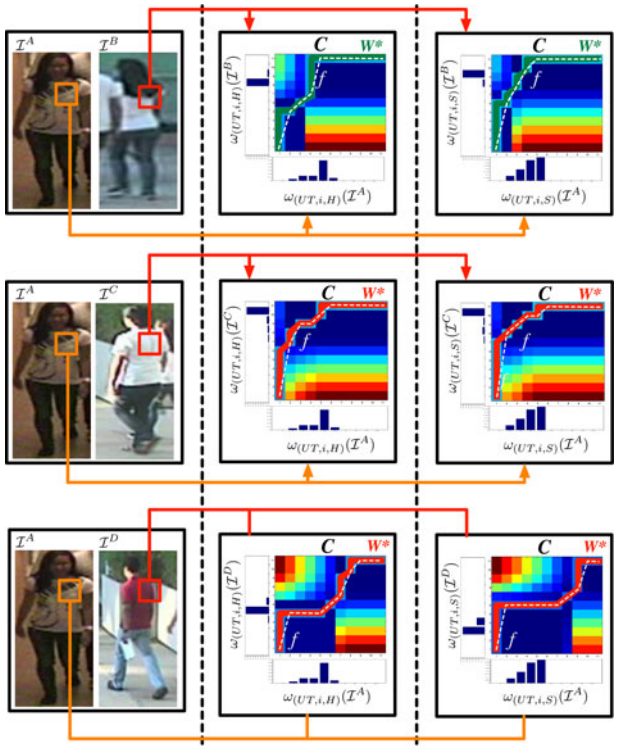

Fig. 7. Example of computing the warp functions between features extracted from the same patch of two images. The first column shows two images from two cameras. The warp function between the features extracted from the same patches (shown by the orange and red boxes) are computed next. The last two columns show the cost matrices, the optimal warp path W^* and the corresponding warp function f. For convenience of visualization, warp functions computed for the H and S colorspaces only are shown in second and third column respectively. The cost matrix is colorcoded and the cost gets higher as the color goes from blue to red. First row shows the feature warps for the same person. Second and third rows show the warping of features between different persons that have similar and different appearance respectively with the person in the left.

$(\omega_{(\phi,i,j)}(\mathcal{I}^A)$ and $\omega_{(\phi,i,j)}(\mathcal{I}^B))$ for each body part ϕ, patch i and feature j. The vector created by concatenating all such vector warp functions computed for the body part ϕ, is denoted as

$$\mathbf{F}_\phi(\mathcal{I}^A,\mathcal{I}^B) = \left\langle \mathbf{f}_{(\omega_{(\phi,i,j)}(\mathcal{I}^A),\omega_{(\phi,i,j)}(\mathcal{I}^B))} \right\rangle, \qquad \forall i,j. \tag{4}$$

The set of all $\mathbf{F}_\phi(\mathcal{I}^A,\mathcal{I}^B)$'s computed between two images $\mathcal{I}^A$ and $\mathcal{I}^B$ of the same person forms the feasible or positive set $\mathcal{F}^p_\phi$ (for bodypart ϕ). The same computed between images of two different persons forms the infeasible or negative set $\mathcal{F}^n_\phi$. Both $\mathcal{F}^p_\phi$ and $\mathcal{F}^n_\phi$ together form the WFS which provides the description of the nonlinear feature transformations under different variabilities.

The proposed WFS model allows us to pose the re-identification problem as finding the parameters of the decision surface, that best separates the sets $\mathcal{F}^p_\phi$ and $\mathcal{F}^n_\phi$. Given a pair of candidate images, we classify such images as coming from the same target or not according as the warp functions between the image features lie in the positive or the negative region.

4.3 Re-Identification in WFS

To re-identify persons moving across camera views we propose to train a binary classifier and classify the warp

TABLE 1
Details and Comparison of Commonly Used Person Re-Identification Benchmark Datasets

Dataset	People	Image info	Cameras	Additional Info
ETHZ [47] (SEQ.#1,SEQ.#2, SEQ.#3)	(83, 35, 28)	Images: (4,856, 1,690, 1,762) Avg. images per person per camera: (59, 48, 63) Size: 13×30 to 158×432	(1,1,1)	Scenario: outdoor Challenges: color changes, occlusions, sptial resolution http://homepages.dcc.ufmg.br/∼william/
CAVIAR4REID [5]	72 (50)	1,220 (1,000) Avg. images per person per camera: 10 (10) Size: 17×39 to 72×144	2	Scenario: indoor Challenges: viewpoint variation, color changes, spatial resolution www.lorisbazzani.info
WARD [8]	70	Images: 4,786 Avg. images per person per camera: 69 Size: 15×36 to 70×189	3	Scenario: outdoor Challenges: viewpoint variations, spatial resolution, color changes http://users.dimi.uniud.it/∼niki.martinel/
VIPeR [48]	632	Images: 1,264 Avg. images per person per camera: 1 Size: 48×128	2	Scenario: outdoor Challenges: viewpoint variation, color changes http://vision.soe.ucsc.edu/node/178
RAiD [1]	43	Images: 6,920 Avg. images per person per camera: 40 Size: 64×128	3	Scenario: outdoor and indoor Challenges: Severe illumination and viewpoint variations, spatial resolution changes http://www.ee.ucr.edu/amitrc/datasets.php

For the CAVIAR4REID dataset, values in brackets are for persons appearing in both cameras. For ETHZ dataset values in brackets are for SEQ.#1, SEQ.#2 and SEQ.#3 respectively.

functions in the WFS as belonging to the feasible or infeasible sets. As discussed in Section 4.2 we use high-dimensional dense color and texture features to represent the appearance of the targets. While it is advantageous for a richer representation, it comes with the curse of dimensionality. The high dimensionality of the features result in high dimensional warp functions. Accordingly, any nonlinear classifier has to pay high computational and memory complexity in the training phase. This scalability issue makes it nontrivial to train a classifier directly on such high dimensional warp functions for large datasets whose training size is typically far beyond thousands. Therefore, we need to select a low dimensional subspace that can adequately handle the intrinsic dimensionality of the warp functions. Towards this objective, and supported by the recent study on real data discussed in [49], we use PCA [35] to embed the WFS into a low dimensional subspace. In the following we refer to $\mathbf{F}'_{\phi}(\mathcal{I}^A, \mathcal{I}^B)$ as the low dimensional warp function computed between images $\mathcal{I}^A$ and $\mathcal{I}^B$ for body part ϕ.

Even though PCA is able to reduce the dimensionality of the WFS, each dimension of it may not, still, be discriminating enough between the feasible and infeasible warp functions. Thus a classifier giving more importance to the more discriminative dimension is desirable. A random forest [36] is a popular and efficient classifier based on bootstrapped aggregation ideas. It is a combination of many binary decision trees built using several bootstrap samples. At each node of each tree a subset of the warp function dimensions is randomly chosen and the best split is calculated only within this subset. This randomization of the warp function dimensions effectively chooses the dimensions according to their importance in separating the feasible and the infeasible warp functions in the WFS. This coupled with the reduction of overfitting error makes RF a suitable choice to learn the parameters of the decision boundary.

In the classification phase the warp function between the features of two candidate images from two different cameras is computed. The trained RF classifies the warp function as coming from the same target or not according as it lies in the positive or the negative region.

Let $\mathcal{I}^{A_1}, \ldots, \mathcal{I}^{A_N}$ be the N images of a given person $\mathcal{A}$ and $\mathcal{I}^{B_1}, \ldots, \mathcal{I}^{B_M}$ be the M images of another person $\mathcal{B}$ in another camera. As commonly accepted in the field of person re-identification, if $N = 1$ and $M = 1$, then the approach is defined to be a *single-shot* approach, otherwise, if both N and M are greater than 1, it is named a *multiple-shot* approach. As the total number of possible warp functions that can be computed for a single body part ϕ is $N \times M$, we have $|\phi| \times N \times M$ predicted probabilities for a target pair, where $|\phi|$ denotes the number of parts into which the body of a person is divided. The probability of $\mathcal{A}$ and $\mathcal{B}$ being the same person is computed by averaging all the $|\phi| \times N \times M$ probabilities obtained from the classifier.

5 Experiments

We evaluated our approach on five publicly available datasets, the ETHZ dataset [52], the CAVIAR4REID dataset [5], the VIPeR dataset [48], the WARD dataset [8] and a dataset (RAiD) [1] collected by us. We chose these datasets because they provide many challenges faced in real world person re-identification applications, e.g., viewpoint, pose and illumination changes, different backgrounds, image resolutions, occlusions, etc. Of these, WARD and RAiD are specifically geared towards large illumination change. More details about each dataset are reported in Table 1 and are discussed below. We report the results for both single-shot ($N = 1$) and multiple-shot ($N > 1$) strategies. For all multiple-shot strategies we use $N = M$. Results are shown in terms of recognition rate as cumulative matching characteristic (CMC) curves and normalized area under curve (nAUC) values, as commonly performed in the literature. The CMC curve is a plot of the recognition percentage versus the ranking score and represents the expectation of

finding the correct match inside top k matches. On the other hand, nAUC gives an overall score of how well a re-identification method performs irrespective of the dataset size. For each dataset the evaluation procedure is repeated 10 times using independent random splits. We reported the average results on these 10 splits. All the results used for comparison were either taken from the corresponding works or by running the publicly available codes on datasets for which reported results could not be obtained. We did not re-implement other methods as it is very difficult to exactly emulate all the implementation details.

5.1 Implementation Details

In our implementation we used the following settings:

- Image pairs of the same or different person(s) in different cameras were randomly picked to compute the positive and negative warp functions respectively;
- $\hat{\mathcal{I}}_{UT}, \hat{\mathcal{I}}_{LT}, \hat{\mathcal{I}}_{UL}$ and $\hat{\mathcal{I}}_{LL}$ have been resized as follows:
  - For the ETHZ dataset: $\hat{\mathcal{I}}_{UT} = \hat{\mathcal{I}}_{LT} = \hat{\mathcal{I}}_{UL} = \hat{\mathcal{I}}_{LL} = 32 \times 16$;
  - For the CAVIAR, WARD and RAiD dataset: $\hat{\mathcal{I}}_{UT} = \hat{\mathcal{I}}_{LT} = \hat{\mathcal{I}}_{UL} = \hat{\mathcal{I}}_{LL} = 64 \times 32$;
  - For the VIPeR dataset: $\hat{\mathcal{I}}_{UT} = \hat{\mathcal{I}}_{LT} = \hat{\mathcal{I}}_{UL} = \hat{\mathcal{I}}_{LL} = 48 \times 32$;
- The size of each dense patch has been selected to be $R \times R = 8 \times 8$ pixels.
- The color histograms extracted from the dense patches were quantized using $b_c = 10$ bins for each color space component c.
- Texture features have been extracted using the following parameters:
  - LBP: we followed the same protocols used in [38]. LBP histograms were quantized into $b_{LBP} = 10$ bins.
  - Gabor: we used Gabor filters at eight orientations and five scales. b_G was set to 16.
  - Schmid: the same filter settings as [40] have been used. b_{Schimd} was set to 16.
  - Leung-Malik: the same filter bank defined in [41] consisting of 36 oriented filters with six orientations, three scales and two phases, eight Laplacian of Gaussian (LoG) filters, and four Gaussians was used. b_{LM} was set to 16.
- δ was taken as the euclidean distance between the feature values.
- While doing PCA, we selected the largest principal components such that the 99 percent of the original variance is retained.[1]
- The RF parameters such as the number of trees, the number of features to consider when looking for the best split, etc. were selected using four-fold cross validation.

The proposed method is, first, evaluated on three challenging benchmark datasets, namely ETHZ, CAVIAR4REID and VIPeR. Since WARD and RAiD contain a large illumination variation, we show the performance on these two datasets separately in the next sub-section.

1. The concatenated feature vector $\mathbf{F}_\phi$ computed for each body part for the RAiD dataset has 55,872 dimensions (i.e. a feature vector of 1,746 dimensions is extracted from each of the $n_\phi = 32$ patches). After applying PCA we obtain $\mathbf{F}'_\phi \in \mathbb{R}^{1223}$. The value is the average number of principle dimensions over all the four body parts and all the 10 trials.

5.2 Comparative Evaluation on Benchmark Datasets

5.2.1 ETHZ Dataset

The ETHZ dataset [52] contains video sequences of urban scenes captured from moving cameras. It contains a large number of different people in uncontrolled conditions. It has originally been proposed for pedestrian detection, but in [47] a modified version of the dataset was provided for the task of person re-identification. This version consists of person images extracted from three video sequences structured as follows: SEQ. #1 containing 83 persons (4,857 images), SEQ. #2 containing 35 persons (1,961 images), and SEQ. #3 containing 28 persons (1,762 images). Since the original video sequences are captured from moving cameras, images have a range of variations in human appearance and some even suffer from heavy occlusions. However, for the same reason the dataset does not provide a realistic scenario for person re-identification with multiple disjoint cameras. To make this dataset more challenging, we followed the strategy proposed in [13] by randomly picking a set of 10 consecutive frames from the beginning and from the end of each sequence.

Despite this limitation it is commonly used for person re-identification, so we also evaluated our approach on this dataset. Following the evaluation setup in [10], [47], all images have been resized to 32×64 pixels. We evaluate our method using both single-shot and multiple-shot strategies. Similar to [19], [25], for the single-shot scenario, we randomly sample two images per person to build a training set, and another two images to build the test set. The test images from one camera constitute the probe and the those from the other camera create the gallery set.

In Table 2 we present the performance of our method using both single-shot and multiple-shot strategies. The first nine rows show the performance comparison with eight different methods when one single image has been used to build the gallery and the probe sets. The last 10 rows show the performance comparison with nine different methods using a multiple-shot strategy. For the single shot scenario our performance is either superior to or same with that of all the eight methods for each of the three sequences. For the multiple-shot scenario the same settings of experiments as in [12], [23] were used with $N = 5$. In this scenario, the BRM [13] approach has superior performances only from rank 1 to rank 4 for SEQ.#1. Similarly the eLDFV [50] method has superior performance compared to our method for rank 1 to 3. Our method is the only one that achieves the 99 percent of correct recognition for this sequence within the top 7 rank scores. On SEQ.#2 we outperform all other methods as we reach 100 percent correct recognition within top 4 matches. Similarly, on SEQ.#3 our method has the best performance and recognizes all the persons at rank 1. Notice that in these experiments we are using $N = 5$ images, whereas the results for SDALF, AHPE, eBiCov and BRM were reported using $N = 10$ images. For all the three sequences in the ETHZ dataset our method is the only one that achieves the 99 percent of correct recognition within the top 7 matches.

TABLE 2
Comparison of the Proposed Method on the ETHZ Dataset Using Both a Single Shot-Strategy (Top 9 Rows) and a Multiple-Shot Strategy (Last 10 Rows)

Method	SEQ.#1							SEQ.#2							SEQ.#3						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Proposed (one image)	**84**	**88**	**91**	**93**	**94**	**95**	**96**	**81**	**86**	**90**	**93**	**95**	**96**	**97**	**91**	**97**	**99**	**99**	**99**	**99**	**100**
eLDFV [50] (one image)	83	87	90	91	92	93	94	79	84	87	90	91	92	93	**91**	94	96	97	97	97	97
SDALF [10] (one image)	65	73	77	79	81	82	84	64	74	79	83	85	87	89	76	83	86	88	90	92	93
eBiCOV [12] (one image)	74	80	83	85	87	88	89	71	79	83	86	88	90	91	82	87	90	92	93	94	95
eSDC_knn [15] (one image)	81	86	89	90	92	93	94	79	84	87	90	91	92	93	90	95	96	97	98	98	99
eSDC_ocsvm [15] (one image)	80	85	88	90	91	92	93	80	**86**	89	91	93	94	95	89	94	96	97	98	98	99
RPLM [25] (one image)	77	83	87	90	91	92	92	65	77	81	82	86	89	90	83	90	92	94	96	96	97
IBML [19] (one image)	78	84	87	89	90	91	91	74	81	84	87	89	91	92	**91**	95	97	98	98	98	99
ICT [34] (one image)	68	76	82	86	87	89	90	70	82	89	91	93	94	95	**91**	94	96	97	97	98	98
Proposed (five images)	94	95	96	97	**98**	**98**	**99**	**98**	**99**	**99**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
PLS [47] (all images)	79	85	86	87	88	89	90	74	79	81	83	84	85	87	77	81	82	84	85	87	89
eBiCOV [12] (five images)	93	94	95	95	96	96	96	91	94	95	96	97	97	97	98	98	99	**100**	**100**	**100**	**100**
eLDFV [50] (five images)	**96**	**97**	**97**	97	**98**	**98**	98	97	98	**99**	**100**	**100**	**100**	**100**	99	**100**	**100**	**100**	**100**	**100**	**100**
LDC [23] (five images)	92	95	96	97	**98**	**98**	98	92	95	97	98	99	99	99	96	97	98	99	99	99	99
ICT [34] (five images)	92	93	94	95	96	96	97	95	98	**99**	99	**100**	**100**	**100**	95	96	97	99	**100**	**100**	**100**
SDALF [10] (10 images)	91	92	93	94	94	94	94	91	94	96	96	97	97	98	94	95	96	96	96	96	96
AHPE [51] (10 images)	85	89	92	93	94	94	95	80	86	89	92	93	94	95	83	91	92	94	96	97	97
eBiCOV [12] (10 images)	93	94	95	96	96	96	96	91	95	96	97	98	99	99	97	98	99	**100**	**100**	**100**	**100**
BRM [13] (10 images)	**96**	**97**	**98**	**98**	**98**	**98**	98	94	95	95	95	95	95	96	96	98	**100**	**100**	**100**	**100**	**100**

Recognition rates for top 7 ranks are shown for each of the three sequences. The best recognition rates for each rank are shown in boldface font.

5.2.2 CAVIAR4REID Dataset [5]

This dataset [5] contains images of pedestrians extracted from the CAVIAR repository. It is composed of 1,220 images of 72 pedestrians out of which 50 are viewed by two disjoint cameras. So, in our approach we considered only these 50 persons. It is more interesting than the ETHZ, where images are extracted from a single camera. Other challenges in this dataset includes a broad change in the image resolution, with a minimum and maximum size of 17×39 and 72×144, respectively, severe pose variations, illumination changes and occlusion.

It is common to split the CAVIAR4REID dataset both in terms of people [24], [34] and not [10], [53]. We conducted experiments following both these protocols to fairly compare against methods following either of these two. Following the same setup as in [34] first, the 50 people are equally divided into training and test sets of 25 persons each. In this setup we compare against LF [24] and ICT [34] who use a multiple shot strategy with $N = 5$ and $N = 10$ images respectively. In Fig. 8a we show that our algorithm outperforms both the methods and reaches as high as 40.9 percent rank 1 score when a multiple shot strategy with $N = 10$ is employed. In the second set up following the same protocol as in [53], we do not split the dataset in terms of persons. Pairs of images are randomly selected in different views for training. The probe and the gallery sets are formed by randomly selecting images from the remaining ones for each person. In this scenario we compare against the methods who have adopted the same strategy of split. Namely the methods are AHPE [51], SDALF [10], CI [14], CPS [5], LAFT [53] and LDC [23]. Fig. 8b shows the CMC curves for the single shot scenario. Figs. 8c and 8d show the comparison with the multi-shot strategy. While for single shot scenario we meet the state-of-the-art performance of LAFT and outperform the rest, for both the multi-shot scenarios we have superior performance over all the compared methods.

5.2.3 VIPeR Dataset

VIPeR [48] is a challenging dataset for person re-identification due to the changes in illumination and pose, and the

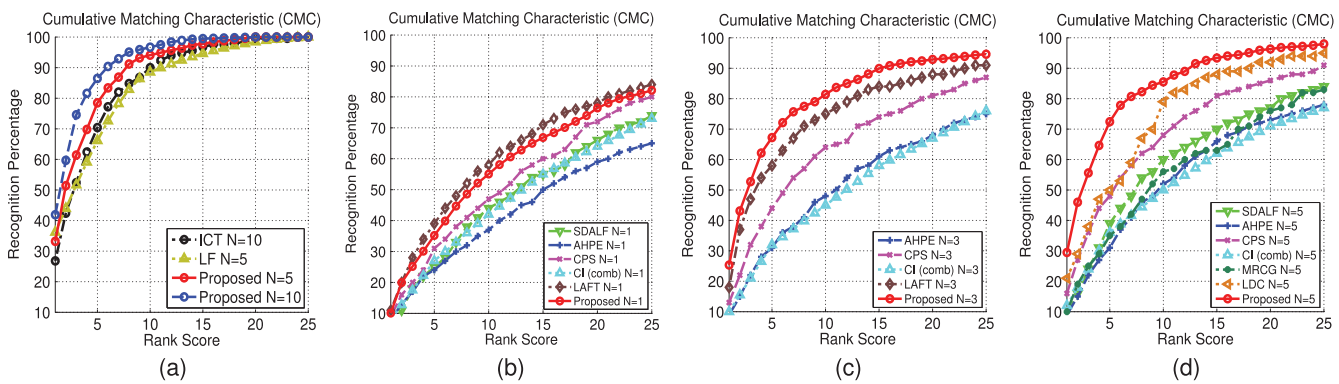

Fig. 8. CMC curves for CAVIAR4REID dataset. In (a) results are shown when the dataset is split in terms of persons. In (b), (c) and (d) comparisons are shown for the case where the dataset is not split in terms of persons with $N = 1$, $N = 3$ and $N = 5$ respectively.

TABLE 3
Comparison of the Proposed Method on the VIPeR Dataset

Rank Score	1	10	20	50	100
Proposed	25.81	69.56	83.67	95.12	98.89
RCCA [16]	**30.00**	**75.00**	**87.00**	96.00	**99.00**
LAFT [53]	29.60	69.30	81.34	**96.80**	**99.00**
LF [24]	24.18	67.12	81.38	94.12	
TML [20]	19.00	61.00	74.00	91.00	97.00
KISSME [22]	19.60	62.20	74.92	91.80	98.00
RPLM [25]	27.00	69.00	83.00	95.00	**99.00**
IBML [19]	22.00	63.00	78.00	93.00	98.00
ELF [54]	12.00	43.00	60.00	81.00	93.00
SDALF [10]	19.87	49.73	65.73	84.80	
PRSVM [55]	14.60	53.90	70.10	85.00	94.00
CPS [5]	21.84	57.21	71.00	88.10	
PRDC [56]	15.70	53.86	70.09	87.00	
LMNN-R [18]	23.70	68.00	80.00	93.00	**99.00**
eBiCOV [12]	20.66	56.18	68.00	84.90	
eLDFV [50]	22.34	60.04	71.00	88.92	**99.00**
eSDC.knn [15]	26.31	58.86	72.77	79.30	
eSDC.ocsvm [15]	26.74	62.37	76.36	82.10	
CI [14]	18.00	50.00	62.00	81.00	
ICT [34]	15.90	57.20	78.30	91.00	95.00
ARLTM [57]	21.00	52.00	68.00	86.00	

Top 100 rank matching rate (percent) is shown.

low spatial resolution of images. This dataset contains one image each from two cameras of 632 persons. Although images from the same camera are not always taken from the same viewpoint and thus do not fully fit our framework, still we compare our results with other methods to show that the proposed approach achieves good results in such a scenario too. To evaluate our method we followed the same normalization approach as in [10], [15], [34], resizing all the images to 48×128 pixels. To compare our approach to state-of-the-art methods we used the same evaluation protocol proposed in [54]. We split the dataset in terms of persons and used 316 of them for training and the remaining 316 for testing. As the VIPeR dataset is a single-shot dataset, we used $N = 1$ images per person to form the training and test sets.

In Table 3 we report the recognition performance for the top 100 ranks and compared the results with 20 state-of-the-art methods for person re-identification. The table shows that the proposed method does achieve a performance better than most of the state-of-the-arts as far as the performance corresponding to rank 1 is considered. It is behind the top performer only by 4.19 percent for rank 1. The performance continuously improves with higher ranks. The rank 100 performance is either same or better than all the methods. According to [34] the performance at higher ranks is, sometimes, more significant as this reflects the algorithm's performance for difficult cases. Thus, in this challenging dataset with only one image per person in two non-static cameras the proposed method does achieve competitive performance as that of the state-of-the-arts.

5.3 Comparative Evaluation with Large Appearance Variation

Since our focus is to understand the space of transformation of features, we provide the performance of the proposed method for two datasets which posses significant appearance variation.

5.3.1 *WARD Dataset*

The WARD dataset [8] contains 4,786 images of 70 different people acquired by three non-overlapping cameras in a real surveillance scenario. This dataset is of particular interest because it has a huge illumination variation apart from resolution and pose changes. We conducted the experiments for all the three different camera pairs, denoted here as camera pairs 1-2, 1-3, and 2-3. The proposed approach is compared with the methods for which either the CMC performance on this dataset is reported in literature or the code is available. Namely the methods are SDALF [10], WACN [8] and ICT [34]. Figs. 9a, 9b and 9c compare the performance adopting a multiple shot strategy with $N = 10$ for camera pairs 1-2, 1-3, and 2-3, respectively. The 70 people in this dataset are equally divided into training and test sets of 35 persons each. For all three camera pairs the proposed method outperforms the rest with rank 1 recognition percentage as high as 51.6 percent for the camera pair 2-3. The next runner up has the recognition percentage of 29.5 percent for rank 1. For all the camera pairs 97 percent recognition performance is reached within top 10 matches.

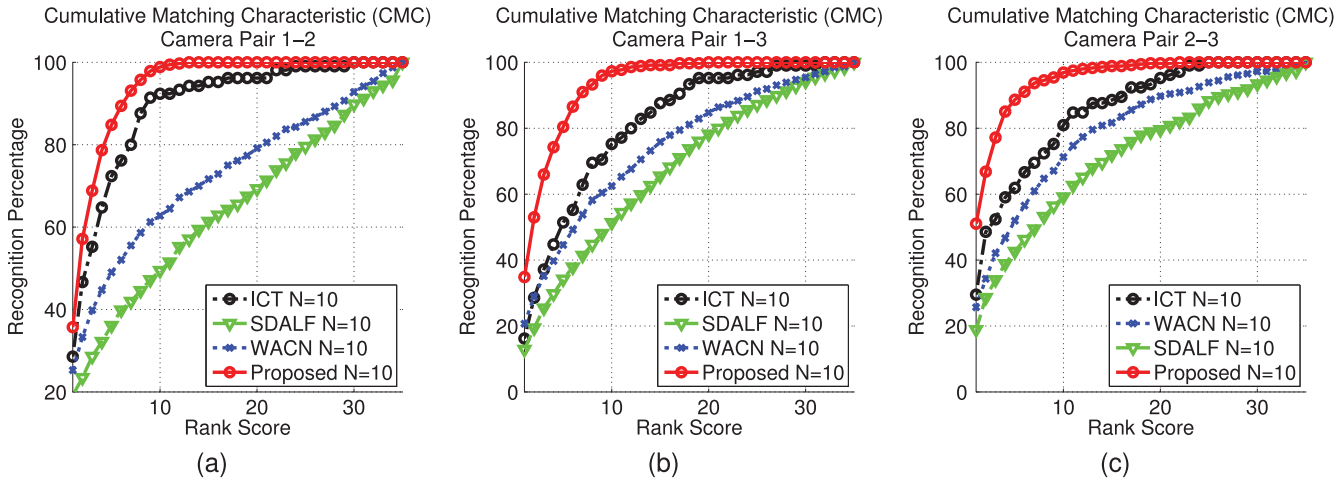

Fig. 9. CMC curves for the WARD dataset. Results and comparisons in (a), (b) and (c) are shown for the camera pairs 1-2, 1-3, and 2-3 respectively. All the results are reported for the case where the dataset is split in terms of persons with $N = 10$.

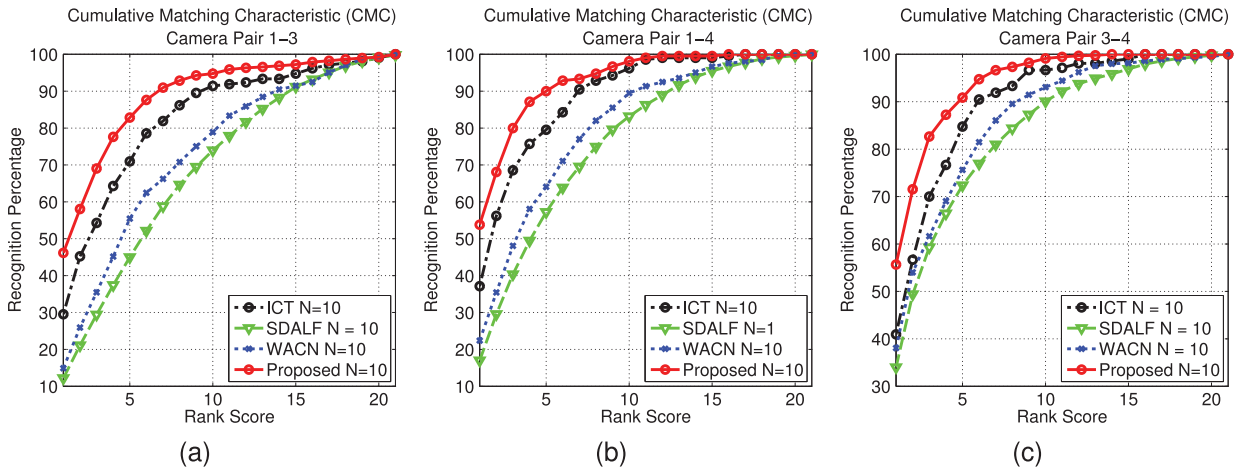

Fig. 10. CMC curves for RAiD dataset. In (a), (b) and (c) comparisons are shown for the camera pairs 1-3, 1-4 and 3-4 respectively.

Fig. 11. Visual comparison of matches using feature warps for camera pair 1-3 of the RAiD dataset. First column is the probe image. Second and third columns show the top 15 matches computed using the proposed method and ICT [34] respectively.

5.3.2 RAiD Dataset

This dataset was collected with a view to have large illumination variation that is not present in most of the publicly available benchmark datasets. In the original dataset 43 subjects were asked to walk through four cameras of which two are outdoor and two are indoor to make sure there is enough variation of appearance between cameras. To reduce the number of pairs of cameras and yet to keep the variation of light to maximum we chose to experiment with three of these cameras, one indoor and two outdoors. These three cameras contain 6,060 images of 41 persons walking through one indoor (denoted as camera 1) and two outdoor cameras (denoted as camera 3 and camera 4). Sample images showing the variation of illumination between the cameras are shown in the supplementary material, available online, and can also be found in the website hosting the dataset.[2]

The proposed approach is compared with the methods for which the code is available. Namely the methods are WACN [8], SDALF [10] and ICT [34]. The dataset was split in terms of persons with 22 persons forming the training set and the rest 21 persons forming the test set. Figs. 10a, 10b and 10c compare the performance adopting a multiple shot strategy with $N = 10$ for camera pairs 1-3, 1-4 and 3-4 respectively. We see that the proposed method is superior to all the rest for both the cases when there is not much appearance variation (camera pair 3-4) and when there is significant lighting variation (for camera pairs 1-3 and 1-4). Expectedly, for camera pair 3-4 the performance is the best achieving 55.7 percent rank 1 performance. For the other two difficult cases too, the proposed method is superior to all the rest achieving 46.4 and 53.9 percent rank 1 performances for camera pairs 1-3 and 1-4 respectively. The second best performance is that of ICT which achieves 29.5 and 37.3 percent rank 1 performances for camera pairs 1-3 and 1-4 respectively. Fig. 11 shows

2. http://www.ee.ucr.edu/~amitrc/datasets.php

TABLE 4
Comparison of Average Performance Across Different Datasets

Number of datasets	4	3				2					
	ETHZ WARD CAVIAR VIPeR	ETHZ WARD CAVIAR	ETHZ WARD VIPeR	ETHZ CAVIAR VIPeR	WARD CAVIAR VIPeR	ETHZ WARD	ETHZ CAVIAR	ETHZ VIPeR	WARD CAVIAR	WARD VIPeR	CAVIAR VIPeR
Proposed	**0.9377**	**0.9292**	**0.9666**	**0.9352**	**0.9200**	**0.9683**	0.9211	**0.9772**	**0.8983**	**0.9544**	**0.9072**
ICT [34]	0.9115	0.8977	0.9302	0.9298	0.8883	0.9188	0.9182	0.9669	0.8561	0.9048	0.9042
SDALF [10]	0.8230	0.7898	0.8538	0.8697	0.7786	0.8195	0.8433	0.9393	0.7066	0.8026	0.8264
RPLM [25]	-	-	-	-	-	-	-	0.9566	-	-	-
IBML [19]	-	-	-	-	-	-	-	0.9549	-	-	-
CPS [5]	-	-	-	0.9062	-	-	0.8914	0.9674	-	-	0.8600
eBiCOV [12]	-	-	-	-	-	-	-	0.9394	-	-	-
eLDFV [50]	-	-	-	-	-	-	-	0.9622	-	-	-
eSDC.knn [15]	-	-	-	-	-	-	-	0.9335	-	-	-
eSDC.ocsvm [15]	-	-	-	-	-	-	-	0.9402	-	-	-
LDC [23]	-	-	-	-	-	-	**0.9250**	-	-	-	-
AHPE [51]	-	-	-	-	-	-	0.8245	-	-	-	-
LAFT [53]	-	-	-	-	-	-	-	-	-	-	0.8820
LF [24]	-	-	-	-	-	-	-	-	-	-	0.8980
CI (comb) [14]	-	-	-	-	-	-	-	-	-	-	0.7948

a comparison of re-identification performances with ICT [34] (achieving the next best performance). The comparison is done on 10 randomly selected persons. For viewing convenience only the top 15 candidates are shown. The green bounding box highlights the ground truth match for each of the query persons. The ground truth match is within top 3 ranked matches for nine out of the 10 examples while six out of these 10 persons are the highest ranked matches too. For the same set of persons the ground truth match is within top 3 ranked matches for two out of the 10 examples in ICT. None of them is the highest ranked match.

5.4 Average Performance across Multiple Datasets

Having shown the performance of the proposed method on separate datasets with different challenges, in this section we show that the proposed method gives the most consistent performance across different datasets each having multiple different challenges. The performance is measured in terms of average nAUC values across different combinations of the four publicly available benchmark datasets (ETHZ, WARD, CAVIAR4REID and VIPeR). We compare with 14 state-of-the-art methods for which either the code is available or results for at least two of these four datasets are reported. The nAUC values for different methods are either taken from the reported results or computed from the reported CMC curves. To make a fair comparison we consider all combinations of two or more datasets and compare our performance by averaging over the datasets separately for each combination. Table 4 shows the performance comparison. The proposed method has the highest average nAUC value for 10 out of the 11 possible combinations. The only case (combination of ETHZ and CAVIAR) where the proposed method is the runner up, the nAUC value changes only at the third decimal place. The superior performance of the proposed method on any combination of these datasets establishes the fact that the proposed method is not tuned to any specific dataset and can address varied number of challenges across different datasets better than the state-of-the-art.

5.5 Robustness to Choice of Classifiers and Patch Size Parameters

To further test the robustness of the proposed method to the choice of classifiers, experiments were conducted with another classifier, namely a support vector machine (SVM) [58]. In a similar way, the proposed method is run with different values of another critical parameter, the patch size of the dense features. We ran these experiments with two datasets, namely WARD and RAiD. In Table 5 we report the recognition performance for different choices of these parameters in terms of the nAUC values. For different choices of the classifiers or for different patch sizes, all the other parameters are chosen as described in Section 5.1. Due to space constraints, we provide the comparison in terms of the CMC curves and the corresponding analysis in the supplementary material, available online. In Table 5 and in the plots provided in the supplementary material, available online, it is shown that the performance is similar even if the classifier is changed to an SVM for both the datasets. As shown in Table 5 the nAUC values differ only at the second decimal places for all the camera pairs with a maximum change of 0.0179 for camera pair 3-4 of the RAiD dataset. Similarly, no major change of the performance is noted for the three different settings of

TABLE 5
Comparison of Performance for Different Choices of Classifiers and Patch Sizes

Dataset	Camera pair	Classifiers		Patch size		
		RF	SVM	4×4	8×8	16×16
WARD	1-2	0.9437	0.9313	0.8996	0.9437	0.9302
	1-3	0.9386	0.9268	0.8896	0.9386	0.9207
	2-3	0.9542	0.9426	0.9081	0.9542	0.9394
RAiD	1-3	0.8905	0.8755	0.8296	0.8905	0.8754
	1-4	0.9295	0.9122	0.8670	0.9295	0.9123
	3-4	0.9395	0.9216	0.8771	0.9395	0.9220

patch sizes for which we conducted the experiment. Indeed the change in the nAUC values is in the second decimal place also for different choice of dense patch sizes with the best performance being observed by a patch size of 8×8. This establishes the robustness of the proposed method to the choice of different classifier types and the dense feature patch sizes.

6 Conclusions

In this work we addressed the problem of multi-camera target re-identification by finding a nonlinear warp function between features from two cameras. Given a pair of feature vectors we show that we can learn the decision surface best separating the feasible and infeasible set of warp functions in the WFS. The target re-identification problem is posed as classifying a test warp function as belonging to the set of feasible or infeasible warp functions. We show that our approach is robust with respect to severe illumination and pose variations by evaluating the performance on five datasets. Our approach outperforms the existing state-of-the-art methods for person re-identification. The future directions of our research will be to apply our approach to capture the transformation of more complex features and to study its application for multi-target tracking in a non-overlapping multi-camera scenario.

Acknowledgments

Amit K. Roy-Chowdhury is the corresponding author. This work was partially supported by the US National Science Foundation (NSF) grants IIS-1316934, CNS-1330110 and Italian Ministry of Defense project "ADVISORIII". The first two authors should be considered as joint first authors.

References

[1] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 330–345.

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

[3] D. J. Bemdt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. Working Notes Knowl. Discovery Databases Workshop*, 1994, pp. 359–370.

[4] M. Müller, "Dynamic time warping," in *Inf. Retrieval for Music and Motion*, vol. 2. Berlin, Germany: Springer, 2007, ch. 4, pp. 69–84.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 68.1–68.11.

[6] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, 2006, pp. 125–136.

[7] F. Porikli and M. Hill, "Inter-camera color calibration using cross-correlation model function," in *Proc. IEEE Int. Conf. Image Proc.*, 2003, pp. 133–136.

[8] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 31–36.

[9] R. Vezzani, D. Baltieri, and R. Cucchiara, "People re-identification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 29:1–29:37, 2014.

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

[11] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. Eur. Conf. Comput. Vis., Workshops Demonstrations*, 2012, pp. 391–401.

[12] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 57.1–57.11.

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

[14] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[15] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3586–3593.

[16] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal-Based Surveill.*, 2013, pp. 244–249.

[17] Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao, "Collaborative sparse approximation for multiple-shot across-camera person re-identification," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 209–214.

[18] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 501–512.

[19] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, 2012, pp. 203–208.

[20] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 31–44.

[21] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2666–2672.

[22] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2288–2295.

[23] G. Zhang, Y. Wang, J. Kato, T. Marutani, and M. Kenji, "Local distance comparison for multiple-shot people re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2013, pp. 677–690.

[24] S. Pedagadi, J. Orwell, and S. Velastin, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3318–3325.

[25] M. Hirzer, P. M. Roth, K. Martin, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 780–793.

[26] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," Michigan State Univ., East Lansing, MI, USA, 2006, https://www.cs.cmu.edu/~liuy/frame_survey_v2.pdf

[27] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *ArXiv e-prints*, 2013.

[28] E. Pkalska and R. P. W. Duin, *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*, vol. 64. Singapore: World Scientific, 2005.

[29] R. Satta, G. Fumera, and F. Roli, "Fast person re-identification based on dissimilarity representations," *Pattern Recognit. Lett.*, vol. 33, no. 14, pp. 1838–1848, Oct. 2012.

[30] R. Satta, G. Fumera, and F. Roli, "A General method for appearance-based people search based on textual queries," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2012, pp. 453–461.

[31] C. Siebler, B. Keni, and R. Stiefelhagen, "Adaptive color transformation for person re-identification in camera networks," in *Proc. Int. Conf. Distrib. Smart Cameras*, Apr. 2010, pp. 199–205.

[32] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," presented at the British Machine Vision Conf., Leeds, U.K., Sep. 2008.

[33] A. Datta, L. M. Brown, R. Feris, and S. Pankanti, "Appearance modeling for person re-identification using weighted brightness transfer functions," in *Proc. Int. Conf. Pattern Recognit.*, 2012, pp. 2367–2370.

[34] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *Proc. Eur. Conf. Comput. Vis., Workshops Demonstrations*, Florence, Italy, 2012, pp. 381–390.

[35] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *J. Educational Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[37] M. Heikklä and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, Apr. 2006.
[38] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[39] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms: Theory and Applications*. New York, NY, USA: Springer, 1998.
[40] C. Schmid, "Constructing models for content-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 2, pp. II–39–II–45.
[41] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, 2001.
[42] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," in *Proc. KDD Workshop Mining Temporal Sequential Data*, pp. 70–40, 2004.
[43] J. Junqua, J. Haton, and H. Wakita, *Robustness in Automatic Speech Recognition—Fundamentals and Applications*. Norwell, MA, USA: Kluwer, 1995.
[44] E. Keogh, "Exact indexing of dynamic time warping," in *Proc. 28th Int. Conf. Very Large Data Bases*, Hong Kong, 2002, pp. 406–417.
[45] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, Dec. 2005.
[46] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa, "Rate-invariant recognition of humans and their activities," *IEEE Trans. Image Proc.*, vol. 18, no. 6, pp. 1326–1339, Jun. 2009.
[47] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. Brazilian Symp. Comput. Graph. Image Process.*, Oct. 2009, pp. 322–329.
[48] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recongnition, reacquisition and tracking," presented at the IEEE International Workshop Performance Evaluation of Tracking Surveillance, Rio De Janeiro, Brazil, Oct. 2007.
[49] L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, vol. 10, pp. 1–41, Feb. 2009.
[50] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Proc. Eur. Conf. Comput. Vis., Workshops Demonstrations*, Florence, Italy, 2012, pp. 413–422.
[51] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by HPE signature," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 1413–1416.
[52] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[53] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3594–3601.
[54] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, Marseille, France, 2008, pp. 262–275.
[55] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 21.1–21.11.
[56] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Jun. 2013.
[57] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, "Attribute-restricted latent topic model for person re-identification," *Pattern Recognit.*, vol. 45, no. 12, pp. 4204–4213, Dec. 2012.
[58] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

Niki Martinel (S'10) received the Laurea degree (cum laude) in multimedia communications from the University of Udine, Italy. Since 2014, he has been working toward the PhD degree in information engineering and he is a member of the AVIRES Lab in the Department of Mathematics and Computer Science at the same university. He has coauthored different scientific works published in international journals and refereed international conferences. He has served as a reviewer for international journals and has been on the organizing and program committees of international conferences. His research interests include machine learning, wide area scene analysis, pattern recognition techniques for surveillance applications, feature transformations and human-computer interaction. He is a student member of the IEEE and IAPR.

Abir Das received the BE and MS degrees in electrical engineering from Jadavpur University, India and the University of California, Riverside, in 2007 and 2013, respectively. He is currently working toward the PhD degree in the Department of Electrical and Computer Engineering, University of California, Riverside. His main research interests include computer vision, person reidentification, multicamera multitarget tracking and video processing using machine learning-based methods. He is a student member of the IEEE.

Christian Micheloni received the MSc and PhD degrees in 2002 and 2006, respectively. He is currently an associate professor at the University of Udine. Since 2000, he has taken part to the European research being under contract for several European Projects. He has coauthored different scientific works published in international journals and refereed international conferences. He has served as a reviewer for several international journals and has been on the organizing and program committees of different international conferences. His main interests involve active vision for the wide area scene understanding, neural networks for the classification and recognition, resource aware camera networks to establish proper control protocols for improving cognition capabilities. He is also interested in pattern recognition and machine learning. His current research projects include camera network self-reconfiguration and person re-identification. He is a member of the International Association of Pattern Recognition (IAPR) and a member of the IEEE.

Amit K. Roy-Chowdhury received the bachelor's degree in electrical engineering from Jadavpur University, Calcutta, India, the master's degree in systems science and automation from the Indian Institute of Science, Bangalore, India, and the PhD degree in electrical engineering from the University of Maryland, College Park. He is a professor of electrical and computer engineering at the University of California, Riverside. His research interests include image processing and analysis, computer vision, pattern recognition, and statistical signal processing. His current research projects include vision networks, distributed visual analysis, wide-area scene understanding, visual recognition and search, videobased biometrics (gait), and biological video analysis. He has authored the monograph *Camera Networks: The Acquisition and Analysis of Videos over Wide Areas*, and published more than 150 papers and book chapters. He has been on the organizing and program committees of multiple conferences and serves on the editorial boards of a number of journals. He is a senior member of the IEEE.