Distributed Signature Fusion for Person Re-Identification

Niki Martinel
University of Udine
Udine, Italy
niki.martinel@uniud.it

Christian Micheloni
University of Udine
Udine, Italy
christian.micheloni@uniud.it

Claudio Piciarelli
University of Udine
Udine, Italy
claudio.piciarelli@uniud.it

***Abstract*—In many surveillance tasks it is very important for security operators to know whether a specific person is present in a given scene, at a given position and time. Person re-identification deals with this problem in order to provide more efficient security. A novel distributed appearance-based method for person re-identification is proposed. Spatio-temporal features are used to group the camera network into camera neighbourhoods. A intra-neighbourhood camera confidence hand-over measure is computed by exploiting a signatures' distance measure. The camera confidence measure is exploited to save network resources. Features that capture the chromatic appearance and the shape of an individual are used to compute a discriminative signature. The Expectation Maximization algorithm is used to fit Gaussian Mixture Models over the chromatic features. GMMs are exploited to compute the distance between signatures and to update the intra-neighbourhood camera confidence. The method has been validated using a benchmark dataset and a new dataset acquired from a wide camera network scenario.**

I. Introduction

Wide area surveillance is gaining a lot interest from the computer vision, sensors and telecommunication communities due to its intrinsic open research issues [12]. As the monitored site grows different problems arise, from the number of sensors to deploy, their configuration, the way they intercommunicate and how they cooperate to achieve a global goal. In this context, even though cameras are becoming cheaper, it is not affordable to have a full coverage of the area, and coverage optimization algorithms should be employed to improve detection probability [13]. Partial area coverage opens to the "blind gaps" problem concerning how to associate objects moving [3] from the Field of View (FoV) of a camera to other ones across not covered zones.

In case of person tracking, re-identification is the way to classify [11] current detections to detections previously achieved by any camera in the network at any location and at any time instant. Re-identification solutions can be categorized into two main categories: biometric, and appearance based methods. The former exploits biometric features and matching techniques to provide across FoV association. The latter relies on the appearance of the objects. The aim of such methods is to extract visual features able to describe a object under different orientations and poses.

In [4], Gallagher and Tsuhan proved that even humans have difficulties in re-identify people without information about clothes. In [14] a color-position histogram descriptor is build on image regions that share similar colors. Bąk *et al.* developed the Mean Riemannian Covariance Grid (MRCG) descriptor [1] by inspecting the features distribution and the appearance temporal changes together with a dense grid structure method. In [7] Hamdoun *et al.* proposed a method that accumulates features from several time-spaced images during tracking. The ensemble of localized features (ELF) approaches by Gray and Tao [6] addresses the problem of viewpoint invariant pedestrian recognition. An AdaBoost algorithm is exploited to learn the signature composed by a combination of spatial and color local features. In [8] a classifier has been trained to learn pairwise dissimilarity profiles between people representations. Doretto *et al.* proposed an algorithm that can be used for generating signatures either from single-images or by accumulating features from multiple images [2].

The proposed work introduces a novel distributed appearance-based method for person re-identification in a wide camera network. Appearance-based features of the same pedestrian acquired by a camera are extracted from multiple frames and accumulated to compute its signature. The Expectation Maximization algorithm is used to fit Gaussian Mixture Models over the chromatic features. A distance measure is defined to compare the distributed signatures built by neighbouring cameras. The distributed approach determines a set of neighbours cameras based on mean pedestrian speed and cameras FoV distances. As the time passes signature matches within the neighbourhood cameras define a confidence hand-over measure. This is used by a camera acquiring a person to ask in a priority way which of the neighbouring cameras previously acquired the current person. This is achieved by sending the current signature to the most confident camera that on the basis of the distance measure answer positively or not. In the first case the re-identification is successful otherwise the next confident camera is checked. An iterative intra-neighbourhood approach is adopted to increase robustness.

Summarizing, the current work introduces the following novelties: i) a novel intra-neighbourhood camera confidence measure for distributed re-identification; ii) the EM algorithm for fitting Gaussian Mixture Models over the chromatic features; iii) a distance measure exploiting the trained GMMs and the other accumulated shape features.

The rest of the paper is organized as follows. Section II gives the system description. The exploited appearance-based

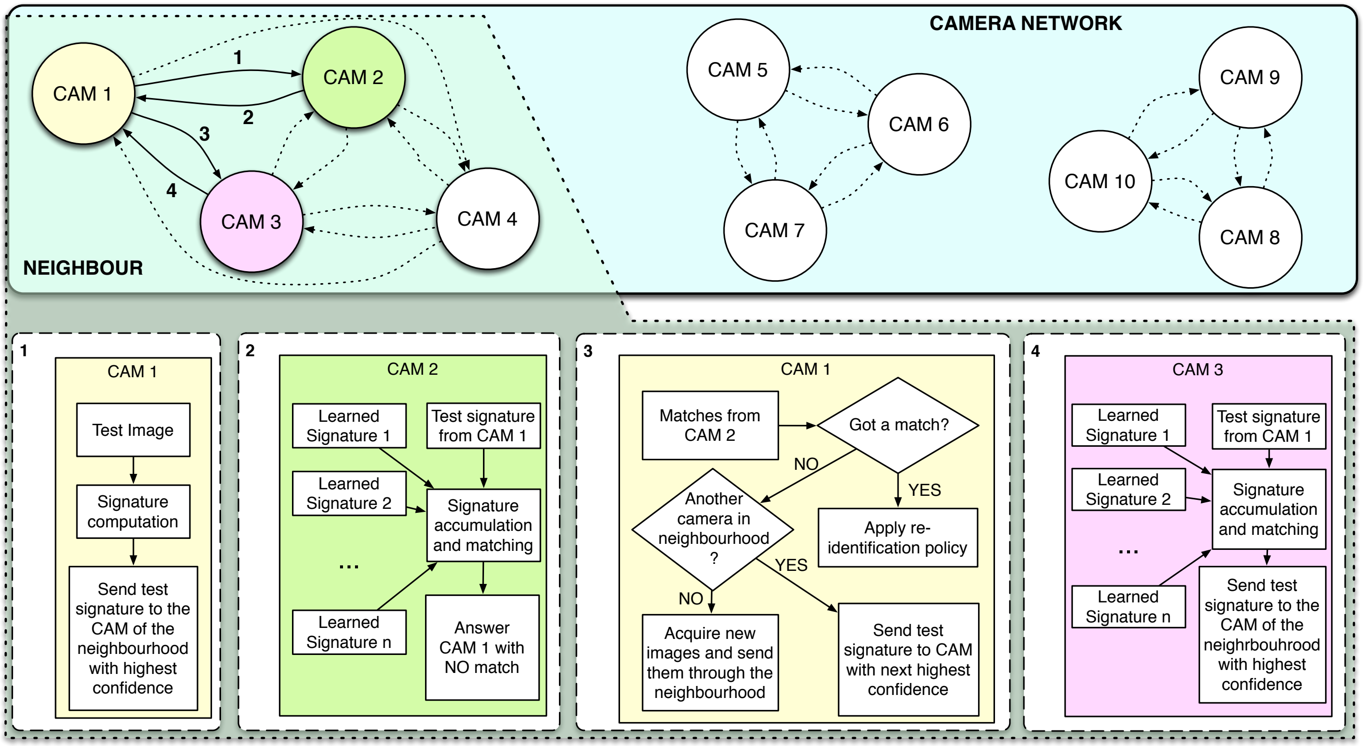

Fig. 1. Distributed re-identification within a camera neighbourhood. The re-identification process within a single neighbourhood is shown.

features are introduced in section III. The methodology for the computation of the proposed signature is given in section IV and the distance measure is described in section V. In section VI the intra-neighbourhood camera confidence measure is described. The distributed re-identification policy is described in section VII. In section VIII experimental results are given. Finally, conclusions are provided in section IX.

II. System Description

The proposed work introduces a novel distributed approach for person re-identification. As shown in Fig. 1, the system computes the intra-neighbourhood re-identification through an iterative process. Network resources are saved exploiting the camera confidence hand-over measure and by using a distributed features approach.

Given a test camera, multiple images of a person are acquired by means of a tracking algorithm. A foreground/background segmentation is used to compute the silhouette of each person further decomposed into the three salient body parts. Finally, three local features are extracted (see section III): i) Pyramid Histogram of Orientation Gradients (PHOG); ii) SIFT; iii) weighted Gaussian color histogram; To provide a pose and orientation invariant signature, an accumulation procedure is introduced to integrate features extracted from consecutive frames.

Given a signature $S_{j,q}$ of the person q built by the camera j, the goal is to determine if the same person has already been detected by other cameras within the neighbourhood $neigh(j)$. A camera confidence hand-off measure is used by camera j to ask in a priority way which of the neighbouring cameras previously acquired the current person. The signature $S_{j,q}$ is initially sent to the most confident camera, let it be k, that computes the distance measure $d(\cdot,\cdot)$ (see eq. 4) between the sent signature $S_{j,q}$ and the signatures S_k computed for all the previously detected people. A match is detected if $d(S_{j,q}, S_{k,t}) < Th_1$ where $S_{k,t}$ is the signature of the t-th person computed with respect to camera k. All the matched signatures are fed back to camera j. If none of the camera k's signatures returned a match for $S_{j,q}$, camera j keeps on analysing new frames of person q in order to update the $S_{j,q}$ signature. For any update of $S_{j,q}$, the described procedure is repeated. To propose a more efficient and distributed approach only the updates of $S_{j,q}$ are sent to cameras that already have received a previous version of the same signature. The procedure is repeated until a valid match is identified or the person q has gone out of the FoV of camera j.

III. Local Features

A feature-based approach is exploited by the proposed method. Three local features are extracted from a given image to capture the appearance of an individual. Such features are accumulated over multiple images by means of a feature accumulation module to compute the signature of a person. A foreground/background separation is initially exploited. The body part division approach exploited in [9] is applied to separate the three silhouette regions B^H, B^T and B^L corresponding to the head, torso and legs respectively. The extracted features and the body part regions are used to compute a discriminative signature by means of the proposed feature accumulation module.

First, PHOG features are computed to capture the shape and the whole appearance of a person. Before extracting such features the given image is projected into the HSV color space to achieve illumination and color invariance. A PHOG matrix $P \in \mathbb{R}^{m \times 3}$ is extracted by concatenating the PHOG histograms extracted from the three image channels. m represents the total number of histogram bins computed by exploiting the original weighted combination of histograms extracted at the different levels of the pyramid representation.

Then SIFT features are computed by exploiting the original cascade filtering approach. The scale space nature of the detector is used to compute the SIFT descriptors. The SIFT features are exploited to capture the chromatic appearance of the individual as proposed in [10].

Finally, for each SIFT feature keypoint vector $p = [x, y]^T$ -where x and y are the coordinates of the keypoint center- a circular image region R of fixed diameter, centred at p, is extracted. Given the region region R, a Gaussian function is used to compute a weighted Gaussian color histogram $H_c \in \mathbb{R}^{b_c}$ for each channel c. The Gaussian function is used such that more weight is given to the part of the region that is less prone to occlusions and pose variations. b_c is the number of bins used for quantization.

IV. Signature Computation

Given the n frames of a person q acquired by camera j, the signature $S_{j,q}$ is defined as $\langle P^{(1,n)}, SIFT^{(1,n)}, H^{(1,n)} \rangle$ where P is the PHOG matrix, $SIFT$ is the SIFT feature vector and H is the weighted Gaussian color histogram. All the three features are accumulated over frames $1, \ldots, n$. The PHOG feature matrix is given by the pooling as $P^{(1,n)} = \frac{1}{n} \sum_{i=1}^{n} P^i$. The SIFT and the weighted Gaussian color histogram features are accumulated exploiting the following three steps: i) match SIFT features; ii) accumulate SIFT features; iii) accumulate weighted Gaussian color histograms of matching SIFT features and update the number of Gaussian distributions used to train a GMM over the weighted Gaussian color histograms.

Let i, l be the match computed exploiting the l^2-norm distance between $SIFT_i^{(1,n-1)}$ and $SIFT_l^{(n,n)}$, the weighted Gaussian color histogram $H_l^{(n,n)}$ is assigned to $SIFT_i^{(1,n-1)}$. Then the distance $d_{\chi^2}(H_i^{(1,n-1)}, H_l^{(n,n)})$ is computed as

$$d_{\chi^2}(H_i^{(1,n-1)}, H_l^{(n,n)}) = \omega(p_i, B_i, p_l, B_l) \cdot \sum_{c=1}^{3} \psi_c \chi^2(H_{i,c}^{(1,n-1)}, H_{l,c}^{(n,n)}) \quad (1)$$

where $H_{i,c}^{(1,n-1)}$ and $H_{l,c}^{(n,n)}$ are the weighted Gaussian color histogram vectors computed for channel c. p_i and p_l are the SIFT keypoints related to the $SIFT_i^{(1,n-1)}$ and $SIFT_l^{(n,n)}$. B_i and B_l are the silhouette regions on which the keypoints p_i and p_l lie. ψ_c is the normalization weight. The function

$$\omega(p_1, B_1, p_2, B_2) = \max(D_M(p_1, B_1), D_M(p_2, B_2)) \quad (2)$$

is used to provide a signature robust to occlusions and pose variations. $D_M(p, B)$ is the Mahalanobis distance between a keypoint p and the silhouette body region $B \in \{B^H, B^T, B^L\}$ onto which the keypoint lie. If the distance $d_{\chi^2}(H_i^{(1,n-1)}, H_l^{(n,n)})$ is higher than a given threshold Th_2 the number of Gaussian distributions that have to be used to train the GMMs over the weighted Gaussian color histograms assigned to $SIFT_i^{(1,n)}$ is incremented by one. All the non-matching features in $SIFT^{(n,n)}$ are accumulated such that $SIFT^{(1,n)} = SIFT^{(1,n-1)} + SIFT^{(n,n)}$. Finally, the Expectation Maximization algorithm is exploited to train the GMMs over all the signature's weighted Gaussian color histogram. Since the feature space over which the GMMs have to be trained is given by number of bins used to quantize each weighted Gaussian color histogram, three GMMs for each SIFT feature are computed.

V. Distance Measure

Given a query signature $S_{j,q}$ of person q computed with respect to camera j and a learned signature $S_{k,t}$ of a person t computed with respect to camera k the signatures distance is given by exploiting a weighted combination of i) the d_{phog} distance between signatures' PHOG features and ii) the d_{wgch} distance between Gaussian weighted color histograms. Here a description about how the distributed signatures are compared is given.

A weighted χ^2 distance metric is exploited to compute the PHOG distance

$$d_{phog}(S_{j,q}, S_{k,t}) = \sum_{c=1}^{3} \kappa_c \chi^2(P^{S_{j,q}}, P^{S_{k,t}}) \quad (3)$$

where $P^{S_{j,q}}$ and $P^{S_{k,t}}$ are the two PHOG matrices. κ_c is the normalization weight.

Let $SIFT^{S_{j,q}}$ and $SIFT^{S_{k,t}}$ be the SIFT features of signatures $S_{j,q}$ and $S_{k,t}$, a match i, l between two SIFT features is given exploiting the l^2-norm distance. Given the match i, l the probability $p(H_i^{S_{j,q}}) = \sum_{g=1}^{K} \lambda_{w_g} \mathcal{N}(H_i^{S_{j,q}} | \lambda_{\mu_g}, \lambda_{\Sigma_g})$ is computed. $H_i^{S_{j,q}}$ is the related weighted Gaussian color histogram. λ_{μ_g} and λ_{Σ_g} are the mean and the covariance of the g-th GMM trained over the weighted Gaussian color histogram assigned to the learned feature $SIFT_l^{S_{k,t}}$. K is the total number of such models. Such a probability is then weighted exploiting eq. (2). Finally, the d_{wgch} distance is given by computing the average probabilities computed with respect to all the matches between $S_{j,q}$ and $S_{k,t}$.

The final distance between a learned signature $S_{k,t}$ and a query signature $S_{j,q}$ is given by

$$d(S_{j,q}, S_{k,t}) = \alpha \, d_{phog}(S_{j,q}, S_{k,t}) + \beta d_{wgch}(S_{j,q}, S_{k,t}) \quad (4)$$

α and β are the normalization weights.

VI. Camera Confidence Measure

A novel camera confidence measure is proposed to achieve a more efficient and distributed re-identification method. The

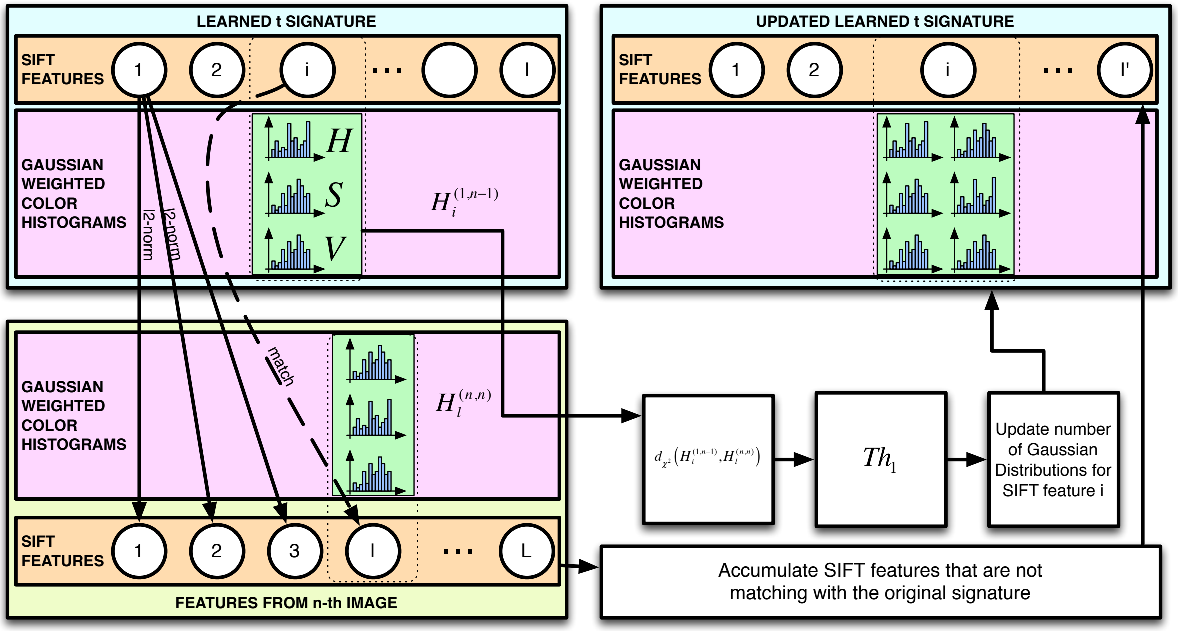

Fig. 2. Accumulation of features. The l^2-norm is exploited to match SIFT features. Weighted Gaussian color histograms of the matched SIFT l are assigned to the learned signature's SIFT feature i. The d_{χ^2} distance between weighted Gaussian color histograms of SIFT feature i and SIFT feature l is computed and a threshold is applied to update the number of Gaussian distribution that will be used to train the GMM. All the non-matching SIFT features are accumulated into a larger signature's SIFT feature vector of length I'.

distance measure between signatures acquired from camera j and all the signatures computed by cameras $\{k|k \in neigh(j)\}$ is exploited to compute the confidence hand-over measure.

The camera confidence measure is learned by exploiting the distance between signatures computed from different cameras. The confidence measure $conf(j,k)$ between all the cameras couples (j,k) in the same neighbourhood C is initially computed during an off-line phase. Then $conf(j,k)$ is updated through the on-line re-identification phase. Given all the signatures computed by j and all the signatures computed by k with $j,k \in C$ the confidence is defined as

$$conf(j,k) = \sum_{q=1}^{Q} \sum_{t=1}^{T} \mathbf{1}_{\{d(S_{j,q},S_{k,t})<Th_1\}} \tag{5}$$

Q and T are the total number of signatures computed for camera j and k respectively.

VII. Distributed Re-identification

As described in section II a query signature $S_{j,q}$ is matched with all the signatures S_k of the same neighbourhood such that $k \in neigh(j)$. The query signature $S_{j,q}$ is initially sent to the most confident camera $k = \arg\max_k conf(j,k)$. Then the distance measure $d(\cdot,\cdot)$ between the $S_{j,q}$ and all the signatures S_k is computed. The process is repeated for all $k \in neigh(j)$ until a match is answered. If no match is answered by k, camera j acquires other frames of person q through a tracking algorithm. Then, given the new frames, the re-identification process is repeated by sending through the neighbours only the features extracted from such images. Those features are used to update the local representation $S_{j,q}$ that has been previously sent to each camera k. The updated signature is compared with the k signatures as described in section V. Since more features are used to compute $S_{j,q}$ a lower threshold $Th'_1 < Th_1$ is exploited. The whole process is repeated until a valid match is identified or the person q has gone out of the FoV of camera j.

VIII. Experimental Results

The proposed method has been validated against a public benchmark dataset and a dataset acquired from a wide area camera network. As suggested in [5] the Cumulative Matching Characteristic (CMC) curve and the Synthetic Recognition Rate (SRR) methods have been used to validate the re-identification method. The true positive rate versus false positive rate curves have also been provided. Each considered datasets has been split into a train set and a test set. The train set is used to compute the initial camera confidence. The test set has been used during the on-line phase to perform the re-identification and to update the camera confidence. N images of the test set are used to compute the signatures of each pedestrian with respect to all the neighbourhood cameras. Evaluation performance have been computed for all the neighbourhood cameras $j \in C$. The match threshold value Th_1 is set to 0.3. If no matches are detected between a query signature $S_{j,q}$ and signatures S_k the proposed features are extracted from W images which are used to update $S_{j,q}$. The reject threshold Th_2 is set to 0.1. The values of α and β have been set to 0.4 and 0.6 respectively. Different values for N and W have been exploited to validate the performance of the proposed method.

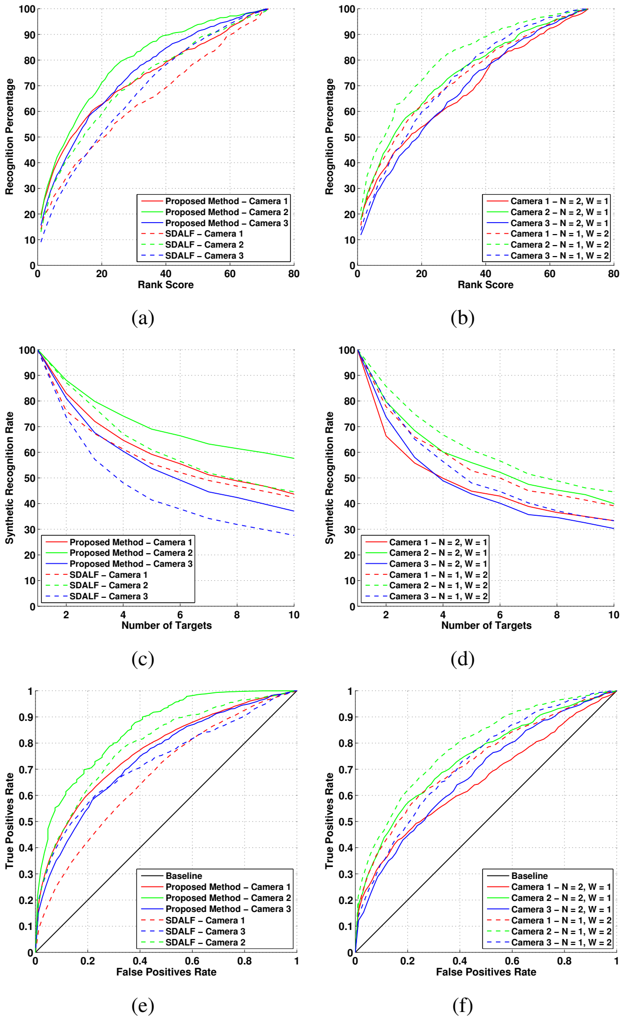

Fig. 3. Performances on the CAVIAR dataset. The re-identification has been performed with respect to each camera. In (a), (c), and (e) results are compared to the method used in [10]. In (b), (d), and (f) the CMC, SRR, and ROC curves shows the performance of the proposed method using different values of N and W.

A. CAVIAR

The CAVIAR4REID re-identification dataset has been used to validate the proposed method. To make a fair comparison with state-of-the-art algorithms [10] images have been normalized to 128×64. The proposed distance is exploited to compare the signatures S_j with the signatures S_k. The camera k is selected by exploiting the proposed confidence measure. The proposed distance provides a ranking for the compared signatures. Evaluation performance have been computed with $N = \{1, 2\}$ and $W = \{1, 2\}$. 100 independent trials for each case have been performed to compute fair results.

In Fig. 3(a), 3(c), and 3(e) the proposed method has been evaluated using $N = 1$ and $W = 1$ images to compute and update the signatures. Results show that the proposed method outperforms the method used for comparison with respect to each considered camera. Even similar performance are achieved by the two method with respect to the rank 1 and rank 2 scores, the proposed method reaches a 70% of correct recognition percentage at the rank 20 scores by considering camera 2. The method used for comparison achieves 59% of correct recognitions at the same rank. The proposed method achieve a 70% of correct recognition percentage at rank 27 and rank 29 by considering camera 3 and camera 1 respectively. As Fig. 3(e) shows the method achieve higher performances and has a true positive rate of about 70% considering a false positive rate of 20%.
The proposed method has also been validated combining different values of N and W. The results of these validations are shown in Fig. 3(b), 3(d), and 3(f). Even similar performance are achieved by the method with respect to the rank 1 and rank 2 scores, by updating the signature with $W = 2$ images many false positive will be rejected and performance increases. A true positive rate of 50% is reached considering a false positive rate of 20% with respect to camera 2 -using $W = 2$ images to update the initial signature. The same true positive rate is reached considering a false positive rate of 31% if $N = 2$ and $W = 1$.

B. Wide Area Re-identification Dataset

To validate the performance of the proposed method with respect to a real surveillance scenario a Wide Area Re-Identification Dataset (WARD) is proposed. 4786 images of 70 different individuals have been captured from three non-overlapping cameras. Since information about the camera displacement and the camera FoVs are available for this dataset, the set of neighbourhood between cameras is computed by exploiting the proposed method. The novel and distributed re-identification approach has been used to compare the signatures S_j with the signatures S_k. The camera k is selected by exploiting the proposed confidence measure. Given a test signature $S_{j,q}$ the proposed distance measure provides a ranking for the signatures S_k. A perfect match is achieved if rank 1 is assigned to the signatures $S_{k,t}$ computed for the same test person with respect to camera k. Performance evaluations have been computed using $N = \{1, 2, 4, 8\}$ and $W = \{1, 6\}$ images to compute the signatures. To make a fair evaluation 100 independent trials have been performed for each case.

In Fig. 4(a), 4(c), and 4(e) the CMC, SRR and ROC curves show the performance of the proposed method with respect to the method used for comparison in [10]. The evaluations have been performed using $N = 1$ and $W = 1$ images to compute and update the signatures. As shown in Fig. 4(a) the proposed method outperforms the method used for comparison with respect to the camera 2 of the neighbourhood. The proposed method achieves a 63% of correct recognitions is reached within the top 10 rank score using the camera 2.
The proposed method has also been validated combining different values of N and W. The results of these validations are shown in Fig. 3(b), 3(d), and 3(f). Since the dataset comes with at least 10 images for each person with respect to each camera the values $N = \{2, 4\}$ and $W = \{2, 4\}$ have been used. Similarly to the CAVIAR evaluations, by using a higher number of images to update the initial signature higher

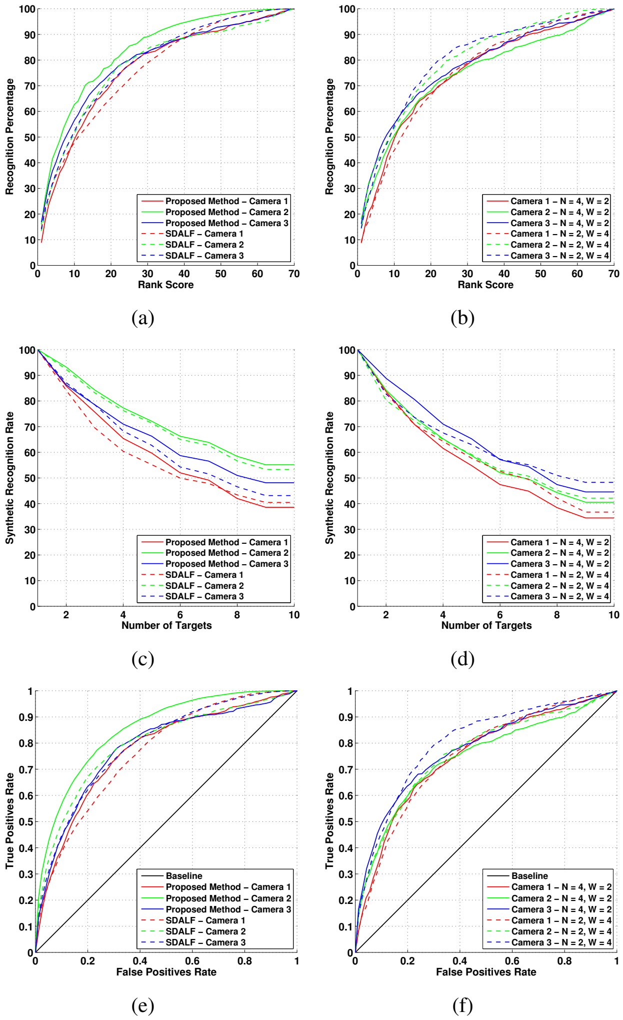

Fig. 4. Performances on the WARD dataset. The three cameras of the neighbourhood have been used to perform the re-identification. Performance of the method with respect to the method also compared in [10] are shown in (a), (c), and (e). In (b), (d), and (f) different values fot N and$W$ have been used.

performance are achieved. A 77% of correct recognitions is reached at the top 20 rank -considering camera 3- by performing the re-identification using $W = 4$ images. Using $W = 2$ images to update the initial signature a 71% of correct recognitions is reached at the same top rank score.

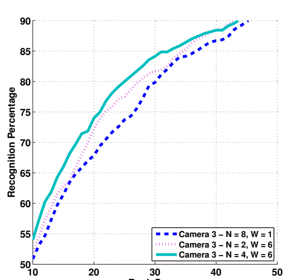

Fig. 5. Evaluations performance with respect to camera 3 is computed selecting different values for N and W.

Fig. 5 shows how the performance of the proposed method increases with respect to each considered camera by using different values for N and W. As it is shown by decreasing N and increasing W higher performance are achieved.

IX. Conclusions

This work introduces a novel distributed appearance-based person re-identification method. Signatures are computed by a temporal accumulation of local features on multiple frames. A camera confidence is exploited to provide a more efficient and distributed signature comparison. A distributed re-identification policy is used to update non-matching signatures by means of a feature fusion method. Evaluation has been proposed using a benchmark dataset and a real surveillance scenario dataset. The proposed method outperforms a state-of-the-art method used for comparisons.

References

[1] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, 2011, pp. 179–184.

[2] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–25, Jan. 2011.

[3] G. L. Foresti, C. Micheloni, and C. Piciarelli, "Detecting moving people in video streams," *Pattern Recognition Letters*, vol. 26, no. 14, pp. 2232–2243, Oct. 2005.

[4] A. C. Gallagher and C. Tsuhan, "Clothing cosegmentation for recognizing people," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage: IEEE, Jun. 2008, pp. 1–8.

[5] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition and tracking," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Rio De Janeiro, Brazil, Oct. 2007.

[6] D. Gray and H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," in *10th European Conference on Computer Vision: Part I*, Marseille, France, 2008, pp. 262–275.

[7] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*. Stanford: Ieee, Sep. 2008, pp. 1–6.

[8] Z. Lin and L. S. Davis, "Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance," in *4th International Symposium on Advances in Visual Computing ISVC'08*, ser. Lecture Notes in Computer Science, vol. 5358, Las Vegas, NV, 2008, pp. 23–34.

[9] N. Martinel and G. Foresti, "Multi-signature based person re-identification," *Electronics Letters*, vol. 48, no. 13, p. 765, 2012.

[10] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Providence, RI: IEEE, Jun. 2012, pp. 31–36.

[11] C. Micheloni, A. Rani, S. Kumar, and G. L. Foresti, "A balanced neural tree for pattern classification," *Neural Networks*, vol. 27, pp. 81–90, 2012.

[12] 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.

[13] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Occlusion-aware multiple camera reconfiguration," in *4th ACM/IEEE International Conference on Distributed Smart Cameras*, Atlanta, GA, USA, 2010, pp. 88–94.

[14] D.-N. Truong Cong, L. Khoudour, C. Achard, and P. Phothisane, "People Re-identification by Means of a Camera Network Using a Graph-based Approach," in *MVA2009 IAPR Conference on Machine Vision Applications*, Yokohama, Japan, May 2009, pp. 152–155.