Abstract

Tracking individuals within a wide area camera network is a tough problem. Obtaining information across uncovered areas is an open issue that person re-identification methods deal with. A novel appearance-based method for person re-identification is proposed. The approach computes a novel discriminative signature by exploiting multiple local features. A novel signature distance measure is given by exploiting a body part division approach. The method has been compared to state-of-the-art methods using a re-identification benchmark dataset. A new dataset acquired from non-overlapping cameras has been built to validate the method against a real wide area camera network scenario. The method has proven to be robust against low resolution images, viewpoint and illumination changes, occlusions and pose variations. Results show that the proposed approach outperforms state-of-the-art methods used for comparison.

1. Introduction

Large camera based monitoring systems have been recently boosted by the eager request of more security. The price plummet of sensors have greatly contributed to the fast growing of wide area camera networks. Despite of this, as the dimensions of a site grow, it quickly becomes hard to deploy a camera network where there are enough overlapping field-of-views to not leave uncovered any portion of the environment. As a matter of fact, in most wide area camera networks there are some areas that cannot be covered by sensors. Such areas, called “blind gaps”, are really critical as it is difficult to obtain any information across them. This particular problem introduces a new issue that is generally referred to as the person re-identification problem. Given the image of a person, the person re-identification problem is to detect if the same person is present in other images acquired by different cameras, at different time instants, or at different locations. This issue is of paramount importance especially to the video analytics systems that cover wide area cameras networks (e.g. airports, metro stations, etc.). Even if in [8] it has been shown that active cameras can be used to track a person within a large area, the proposed solution cannot be always provided.

The proposed person re-identification method exploits four local features to compute a discriminative signature for each person. The method applies to both the single-shot and multiple-shot cases, and it deals with sensors’ low resolution, pose variations, illumination changes, etc. The method is independent from the number of considered candidates and can deal with an arbitrary large number of individuals. Five different modules are proposed to perform the re-identification: i) background/foreground segmentation; ii) body parts division; iii) local features computation; iv) novel discriminative signature computation; v) novel distance measure evaluation. The proposed distance measure is evaluated against each body part so that the effects of occlusions, and pose variations are minimized. The main novel aspect of the proposed algorithm is given by the way the proposed features, together with the novel signature computation, are jointly used to perform the person re-identification task.

The proposed approach introduces the following novelities: i) no training is needed on the set of candidates. Each new individual image is inspected individually; ii) a novel signature that deals with pose variation, occlusion, illumination changes is proposed; iii) the detected body parts are exploited to compute a novel distance metric; iv) novel distance measure evaluation. The proposed work has been evaluated using the i-LIDS re-identification dataset. To evaluate the performance of the method with respect to a wide area camera network, a new dataset has been acquired from a real scenario.

The rest of the paper is organized as follows. Related works are described in section 2. The system description is given in section 3. Here the main modules of the proposed method are introduced. In section 4 computation of local features is described. The novel distance measure is introduced in section 5. In section 6 experimental results are given. Finally, conclusions are provided in section 7.
2. Related work

Within the re-identification community two main groups of methods have been identified. Techniques that deal with visual information and methods that exploit biometrics features. Methods of the former group are commonly referred to as appearance-based methods. Methods of the latter group exploit biometrics features. These approaches are named biometrics-based person re-identification methods.

Appearance-based methods use only visual information to build a person’s signature. Appearance-based methods are organized into two main approaches: i) single-shot methods and ii) multiple-shot methods. Methods that belong to the former group compute the signatures by using a single image for each person; Methods of the latter group use more that one image of the same person to build a single signature. A multiple set of images of the same person is considered as training data, another set is taken as test data.

Single-shot methods use to divide individual’s body before extracting local features. Wang et al. have proposed a single-shot method [10] that exploits two layers to extract features from the identified body parts. A body part segmentation has also been proposed in [5]. A weighted combination of three different local features is evaluated against the symmetry and asymmetry body axes.

As for single-shot approaches, multiple-shot approaches also use to segment individuals’ body. In [1] the features distribution and the appearance temporal changes together with a dense grid structure method have been exploited to develop the Mean Riemannian Covariance Grid (MRCG). Gandhi and Trivedi [6] have exploited camera tracking and camera geometry to develop the Panoramic Appearance Map (PAM). Doretto et al. [4] have proposed an algorithm that can be used for generating either single-shot or multiple-shots signatures.

In contrast with appearance-based methods, biometrics-based methods exploit passive biometrics in order to build an individual signature. Despite the efforts in the field of face, and gait [2] recognition, passive biometric features generally need a precise camera configuration and sensor deployment.


3. System description

As shown in Figure 1, the proposed system introduces a cascade approach to build the novel discriminative signature. The first module applies a foreground/background separation. This is required to provide a fair comparison with other state-of-the-art methods. The second module of the system separates the three salient body parts. The approach proposed in [5] is used. Then the signature computation module exploits the histogram equalized silhouette to extract the local features. The computed features are used to build the novel discriminative signature. Finally, given the signature of a probe image and the signature of a gallery image, the novel distance measure is computed.

4. Features and signature computation

The proposed method extract three local features from each image. These features are used to build a novel discriminative signature. SIFT features are computed by exploiting the proposed cascade filtering approach. SIFT features are used to capture the chromatic appearance of an individual. In order to deal with illumination changes and color variations each silhouette is projected into the HSV color space before computing the weighted Gaussian color histogram features. For each SIFT feature keypoint vector $k = [x, y]^T$ -where $x$ and $y$ are the coordinates of the keypoint center- a circular image region $S$ of diameter $\phi$,
Figure 2. Computed features: (a) weighted Gaussian color histogram features are computed by extracting three histograms from a circular image region centred at the same location of the detected SIFT feature keypoint. (b) PHOG features are computed for each image channel. Here the PHOG extraction process for the first image channel is represented. (c) two Haralick feature vectors are computed for the two main body part regions T and L.

centred at $k$, is extracted. A Gaussian distribution function is used to compute the weighted Gaussian color histogram feature vector $H_t \in \mathbb{R}^{h_t}$ over each region $S$, $b_i$ and $i$ represent the number of bins used for quantization and the image channel from which the histogram is extracted, respectively.

To compute the Pyramid of Histograms of Orientation Gradients (PHOG) features the silhouette is projected into the HSV color space. Then, three PHOG feature vectors $phog_i \in \mathbb{R}^m$ are extracted. $m$ and $i$ are the number of histogram bins and the image channel from which the histogram is extracted, respectively. PHOG features are finally accumulated in $PHOG \in \mathbb{R}^{m \times I}$. $I$ is the total number of the image channels.

As Figure 2(c) shows, the Haralick texture features are extracted from the two detected body part regions $T$ and $L$. Since these features are based on the Gray Level Co-occurrence Matrix (GLCM), the considered body region are converted to the gray scale color space. Intensity values of both regions $T$ and $L$ are scaled to $N_g$ gray levels to reduce computational costs. As shown in Figure 2(c), four directional co-occurrence matrices are computed using four directions of adjacency. The four GLCM are finally used to extract the Haralick features $f_1, f_2, \ldots, f_{14}$. Thus, for each signature two feature vectors $harr_T \in \mathbb{R}^{14}$ and $harr_L \in \mathbb{R}^{14}$ are computed. Finally, all the computed features are used to build the novel discriminative signatures.

In the multiple-shot cases the signature of a given person is computed by accumulating features from multiple images. Given multiple images of a person $A$, the weighted Gaussian color histogram features, and SIFT features are accumulated to build a single signature. The PHOG feature matrix $PHOG$ is computed by averaging the PHOG matrices extracted from the given images. Similarly, the Haralick feature vectors from multiple images are fused by computing the mean vectors.

5. Distance measure

The proposed method computes a novel distance measure by exploiting a weighted combination of the distance between the weighted Gaussian color histograms features, the PHOG features, and the Haralick features.

Before evaluating the weighted Gaussian color histogram distance, two steps are computed. The $l^2$-norm distance is computed to detect the SIFT matches. A RANSAC approach and the body part division are exploited to detect outliers. The body part division is used such that, if two matching SIFT feature keypoints do not lie on the same body part the match is rejected. These improves the results because SIFT features that do not belong to the same body part are not considered within the weighted Gaussian color histogram distance computation. Given a SIFT match, a $\chi^2$ weighted distance measure is employed to match the related weighted Gaussian color histogram features. The weighted $\chi^2$ distance is computed as

$$d_{\chi^2}(A_H, B_H, A_k, B_k) = m(A_k, B_k) \sum_i \psi_i \chi^2(A_{H_i}, B_{H_i})$$

where $A_H$ and $B_H$ are the weighted Gaussian color histogram vectors computed for the signature $A$ and the signature $B$ on channel $i$. $A_k$ and $B_k$ are the two matching SIFT feature keypoints. The $\chi^2(\cdot, \cdot)$ function computes the $\chi^2$ distance metric and $\psi_i$ is the normalization weight. The weighting function $m(\cdot, \cdot)$ exploits a Mahalanobis distance.
metric in order to deal with occlusions and pose variation. Thus each histogram distance is weighted by
\[ m(k_1, k_2) = \max(d_M(k_{1x}, k_{1y}), d_M(k_{2x}, k_{2y})) \]
where \( k_1 \) and \( k_2 \) are two SIFT feature keypoint vectors, and \( d_M(\cdot, \cdot) \) is the Mahalanobis distance metric computed for each body part of the detected silhouette. Finally, given the signature \( A \) and the signature \( B \), the overall weighted Gaussian color histogram distance \( d_{wgh}(A, B) \) is given by the mean \( d_{\chi^2} \) distance computed for each detected match.

Similarly to the computation of the weighted Gaussian color histogram distance, a weighted \( \chi^2 \) distance metric is used to compute the distance between PHOG features. Given the PHOG feature matrices of two signatures \( A \) and \( B \), the PHOG distance is computed as
\[ d_{phog}(A, B) = \sum_i \lambda_i \chi^2(A_{PHOG_i}, B_{PHOG_i}) \]
where \( A_{PHOG_i} \) and \( B_{PHOG_i} \) are the PHOG feature vectors computed for the signature \( A \) and \( B \) on channel \( i \). \( \lambda_i \) is the normalization weight.

Given two signatures \( A \) and \( B \) a \( l^2 \)-norm distance is used to compare the Haralick feature vectors that are computed over the same body part. The final distance \( d_{har}(A, B) \) is given by the mean \( l^2 \)-norm distance computed for each pair of Haralick feature vectors.

Finally, the overall distance between two image signatures \( A \) and \( B \) is given by
\[ d(A, B) = \alpha d_{wgh}(A, B) + \beta d_{phog}(A, B) + \gamma d_{har}(A, B) \]
where \( \alpha, \beta, \) and \( \gamma \) are the normalization weights.

6. Experimental results

The proposed method has been compared with state-of-the-art methods using a benchmark dataset and a dataset acquired from a wide area camera network. Each dataset covers different aspects and challenges posed by the person re-identification problem. As suggested in [7], the Cumulative Matching Characteristic (CMC) curve and the Synthetic Recognition Rate (SRR) methods have been used to validate the proposed method.

Given a gallery set \( A \) and a test set \( B \), to validate the method, three different approaches have been used: i) single-shot versus single-shot (SvsS) method: each given image represents a different person; ii) multiple-shot versus single-shot (MvsS) method: a single person in the set \( A \) is described by a signature made of multiple images’ features. Each image in the set \( B \) represents a different individual; iii) multiple-shot versus multiple-shot (MvsM) method: both signatures of individuals of set \( A \) and \( B \) are made by accumulating features over multiple images.

\[ \alpha \text{Rank Score} \]
\[ \text{Number of targets} \]
\[ \text{Synthetic Recognition Rate (SRR)} \]

Figure 3. Performances on i-LIDS dataset. In (a) and (b) the proposed method is compared to [5] and [11]. According to [5] only the first 25 rank position are displayed on the CMC curve plot. (c) and (d) represent the CMC and SRR curves computed to validate the method against both the MvsS and the MvsM case. Different values of \( N \) have been used.

6.1. i-LIDS dataset

The i-LIDS dataset has been captured in a real environment. The 479 images of 119 individuals images have been acquired from non-overlapping cameras under meaningful illumination changes and occlusions. Both single-shot and multiple-shot cases have been evaluated. The method has been compared with [5] and [11]. To make a fair comparison, images have been normalized to 128×64 and no additional information about the context has been added. In the single-shot case, a single image for each individual has been randomly selected to build the gallery and the probe set. Considering images from the gallery set and images from the probe set, the novel distance is exploited to compare each image of the probe set with the images of the gallery set. The given distance provides a ranking for every image in the gallery with respect to the probe. Ideally rank 1 should be assigned only to the correct pair matches. 10 independent trials have been performed to fairly estimate the performance of the method. As Figure 3(a) and 3(b) show the proposed method outperforms both [5] and [11]. The method get up to 27% rank 1 correct matches.

As to multiple-shot, both MvsS and MvsM cases have been evaluated. For the former, \( N \) images of the same person have been used to build a single signature in the gallery set. A single image for each person is used to built the signatures of the probe set. For the latter, \( N \) images of each pedestrian have been used to build a single signature in both
the gallery and the probe sets. In both cases, the \( N \) images of the same person have been randomly selected. Since the dataset has at least four images for each person, the method has been evaluated with \( N=2 \) and \( N=3 \) for the MvsS case, and \( N=2 \) for the MvsM case. 100 independent trials have been run for each case. As Figure 3(c) and Figure 3(d) show, the method get up to 40% rank 1 correct matches by using just \( N=2 \) images for each person in the MvsS case. If another image is used to compute the multiple-shot signature, an increment of 22% -with respect to the single-shot case- is reached. Similar performance is achieved using \( N=2 \) images to compute both the signatures of the gallery set and the test set in the MvsM case. A 50% rank 1 correct matches is reached.

In Figure 4 the top 5 ranks computed by [5] and the proposed method are depicted. The given results are computed by considering the SvsS case. According to results shown in Figure 3(c) the method proves to be robust enough to deal with occlusions and pose variation. The most erroneous matchings are due to severe illumination changes, or occlusions.

6.2. Wide area scenario dataset

To validate the method against against a wide area camera network a new dataset has been extracted from a real scenario. As Figure 5 shows images have been acquired by three non-overlapping cameras. The three uncalibrated sensors acquire images with a spatial resolution of 320×240 at 1fps. About 3 images for each person have been extracted from CAM A, CAM B and CAM C. A total of 196 images of 61 different individuals has been extracted. Images have been selected such that pose variations, illumination changes and occlusions are maximized. Large changes in resolution are also covered by the dataset, i.e., sizes vary from 36×15 to 189×70. Images have been normalized to 96×48. The dataset has been split into three sets: CAM A, CAM B, and CAM C.

As for the i-LIDS dataset, experiments on both the single-shot and multiple-shot cases have been performed. For the former case, two images of the same person have been randomly selected from two different camera sets. The two camera sets have also been randomly chosen. The selected images have been used to build the gallery and the probe set. The novel distance measure is exploited to compare the gallery set and the test set. The novel distance measure is exploited to compare the gallery set and the test set. For each person in the probe set the position of the correct match is obtained.

In Figure 6(a) and Figure 6(b) results of comparisons with [5] are given. The single-shot case is used. It can be seen that the proposed method outperforms [5]. In particular, the method gets up to a 22% rank 1 matching rate, versus around 12% in [5]. The most erroneous matchings are due to the large viewpoint variations and to the low resolution of images.

Figure 6(c) and Figure 6(d) show the comparative graphs for the CMC and SRR curves. Both MvsS and MvsM tests have been performed. For the former, a gallery set of multiple-shot signatures is used to match a probe set made of single-shot signatures. For the latter, multiple-shot signa-
tures have been used to built both the gallery and the probe set. In both cases, the multiple-shot signatures are made up with $N$ randomly selected images of the same person. Since the dataset has a mean of about 9 images for each person, the method has been evaluated with $N=2, 3, 5$ for the MvsS case, and $N=2, 3$ for the MvsM case. 100 independent trials for each case have been performed. In case of MvsS evaluations the method gets up to 37% rank 1 correct matches by using just $N=2$ images for each person. Up to a 45% rank 1 correct matches is reached both using $N=3$ images for the MvsS case and $N=2$ images for the MvsM case. As Figure 6(c) and Figure 6(d) show, performance of the method increases up to to a 64% rank 1 correct matches by considering more images to build the novel signature. In particular using $N=5$ images for the MvsS case and $N=3$ for the MvsM case an increment of 36% and of 41% -with respect to the single-shot case- is reached, respectively.

7. Conclusions

This work introduces a novel appearance-based person re-identification method. The person re-identification problem has been addressed by computing four local features that capture the shape, the chromatic content and the texture information of an individual. The extracted features are used to build a novel discriminative signature. A novel weighted distance measure has been defined to compare the signatures. The distance measure evaluates the local features extracted from the individual’s body parts in order to prune non reliable matches. The method has been compared with state-of-the-art methods exploiting both the single-shot and the multiple-shot approaches. A benchmark public dataset and a wide area scenario dataset have been used to validate the method. Results show that the method outperforms state-of-the-art methods used for comparison.

References