1. ABSTRACT

Person re-identification through a camera network deals with finding a correct link between consecutive observations of the same target among different cameras in order to choose the most probable correspondence among a set of possible matches. This task is particularly challenging in presence of low-resolution camera networks. In this work, a method for people re-identification in a framework of low-resolution camera network is presented. The proposed approach can be divided in two parts. First, the illumination changes of a target while crossing the network is analyzed. The color structure is evaluated using a novel color descriptor, the Color Structure Descriptor, which describes the differences of dominant colors between two regions of interest. Afterwards, a new pruning system for the links, the Target Color Structure is proposed. Results shows that the improvements achieved applying Target Color Structure control are up to 4% for the top rank and up to 16% considering the first eleven more similar candidates.

Keywords: Video Surveillance, Person Identification, Context, Histograms, Image color analysis

2. INTRODUCTION

Video surveillance is a common tool used for protecting people and property. Currently, due to the advances in technology and the increasing availability at low prices of cameras, the creation of a video set up is not expensive and this phenomenon has led to the use of video surveillance systems also in low-risk environments as private buildings, shops, and residential neighborhoods.

An open issue in video surveillance is the definition of an automatic system able to substitute a human observer in charge of analyzing the recorded video content. In particular, an important problem is person re-identification that is, given the image of a person in one camera, to track him/her over the network by finding instances of the same person in videos recorded by different cameras. The most popular approach used to deal with this problem is the usage of appearance-based algorithms that take into account colors and textures. Other systems are based on identifying targets based on motion or trajectory estimation. However, appearance-based algorithms are sensitive to variations in camera angle, illumination conditions, pose, and appearance of clothes. Moreover, since an object can have a large number of potential matches, a useful application that can be used for speeding up the identification procedure, is the automatic selection of a set of most probable matches to be proposed to the human supervisor.

The proposed approach defines a method for target re-identification and it is based on two steps. First, the illumination changes affecting the camera network are studied. In order to do that, the transformation changes in the dominant colors worn by targets while they move from the field of view of a camera to another, is acquired and classified into one of five linear illumination change models. The color structure is evaluated by means of a novel color descriptor called CSD (Color Structure Descriptor) which has the property of being intensity, scale, and shift invariant. The CSD allows to describe the difference of dominant colors between two regions of interest, in this case the shirt and the pants of the target. Due to the property of invariance of this descriptor, the distance between two CSDs evaluated on the same target viewed from two different cameras is up to 4% for the top rank and up to 16% considering the first eleven more similar candidates.
very low. For this reason, the link between objects is considered only if the distance between two observations in term of CSD is inferior than a threshold. Then, a new pruning system called TCS (Target Color Structure) is proposed: the link between two observations is taken into account if and only if there is coherence in the color structure worn by targets.

In the rest of the paper more details on the proposed approach will be given. In particular related works are introduced in Section 3. A formulation for the multi-camera tracking system is given in Section 4 and a model for the illumination changes over the network is defined in Section 5. Section 6 describes the CSD and Section 7 details the usage of TCS control. Section 8 presents the achieved results and finally the conclusions are drawn in Section 9.

3. RELATED WORK

As mentioned in the introduction, re-identification methods relying on visual information are addressed to as appearance-based techniques. To this group belong methods based both on soft biometrics and on global appearance. The first group of methods has high quality constraints so we will focus our attention on the second one.

In order to handle the modification in appearance due to illumination changes, the global appearance can be evaluated using illumination invariant color descriptors (1, 2, 5, 6, 8, 9, 13, 14, 21) or computing the matching in a transformed sub-space, where the transformation is given by learning the Brightness Transfer Function (BTF) between cameras (12, 16, 18, 26, 27). Other approaches assume easier operative conditions: they simplify the problem by adding temporal reasoning on the spatial layout of the monitored environment, in order to prune the candidate set to be matched (11, 15, 17-19, 22-25). In the inter-camera Brightness Transfer Function is evaluated for multi-camera tracking purposes; probabilistic PCA (Principal Component Analysis) is used to calculate the subspace of BTFs for a set of known correspondences. However, this method relies on training subjects with a large range of brightness values in order to have an accurate BTF.

Prosser et al in 27 use a BTF-based method but in this case the BTF is computed after the collection of training data; the same authors in 26 propose a novel method for multi-camera people association based on adapting Cumulative Brightness Transfer Function (CBTF) to new illumination conditions without the need for a manual training stage using new foreground objects. Also Gilbert and Bowden in 12 moved in this direction modeling inter-camera colour transformations using an incrementally updated transformation matrix.

The approach presented in this paper is based on the classification of illumination changes in one of the five models as proposed in 28; to the best of our knowledge these models, used for image retrieval, have never been used for video surveillance purposes. After the illumination change is classified, a descriptor which is invariant with respect to illumination changes is defined and it is used to prune the number of possible matches for each target to be re-identified.

4. FORMULATION OF MULTI-CAMERA TRACKING PROBLEM

Let us consider a system composed of M cameras C1, C2, ..., CM. Each camera view is divided in l regions so that the whole network is divided in n = m * l regions; that is to say R = {Rk|k = 1 : n}. Each camera sees v targets p so let us indicate all targets seen by the generic camera Ck as pk = p1,k, ..., p,v,k. Each target generates an observation. Let O be the set of all observations performed in the camera network and let Oi be the set of observations performed by the camera Ci on the target pi

\[ O_i = O_{i,1}, ..., O_{i,v}. \]  

The problem of tracking object in a multi-camera network, lies in linking observations of the same object that belong to different cameras. It is then possible to consider the observations of each object as a chain of observations 18, 19. In this way, our problem turns in the search of a correct linkage between consecutive observations, as shown in Figure 1.

Assuming that the single camera tracking problem has already been solved, the multi-camera tracking task basically consists in linking the observations of an object exiting the field of view of one camera to its
observations when entering the field of view of another camera. The optimum solution to this task is obtained by solving a ML (Maximum Likelihood) problem\textsuperscript{18}. Let an hypothesized correspondence between two consecutive observations, \(O_i;O_j\), be denoted as \(k_{i,j}\). Let \(\Phi_{k_{i,j}}\) be a binary random variable which is true if and only if \(k_{i,j}\) is a valid hypothesis, i.e., \(O_i;O_j\) are consecutive observations of the same object. Our task is to find a set of correspondences \(K = \{k_{i,j}\}\) in order that \(k_{i,j} \in K\) if and only if \(\Phi_{k_{i,j}}\) is true.

The proposed solution is based on modeling the system by using the graph theory. In particular, we consider a directed graph in which the hypothesized correspondence \(k_{i,j}\) is represented by an arc from the vertex of observation \(O_i\) to the vertex of observation \(O_j\). In this way it is possible to obtain a concatenation of trees in which a set of observations belonging to the set \(O_i\) are the parent nodes, i.e. observations on the targets need to be re-identified, and the observations belonging to the set \(O_j\) are the child nodes, i.e. observations belong to the candidates.

5. LEARN ILLUMINATION CHANGES IN THE CAMERA NETWORK

Changes in illumination can be modeled by diagonal offset mapping model\textsuperscript{10} as follows:

\[
\begin{bmatrix}
\tilde{R} \\
\tilde{G} \\
\tilde{B}
\end{bmatrix} =
\begin{bmatrix}
a & 0 & 0 \\
0 & b & 0 \\
0 & 0 & c
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
+ \begin{bmatrix}
o_1 \\
o_2 \\
o_3
\end{bmatrix}
\]

in which the colors that are taken into account under an unknown light source (R,G,B) are mapped to their corresponding colors under the canonical illuminant \((\tilde{R},\tilde{G},\tilde{B})\). Moreover an offset \(o_i\) is introduced in order to take into account the "diffuse" light term.

Based on the diagonal model and the diagonal offset model, it is possible to define five different classes of illumination changes:

1. light intensity change: image values change by a constant factor in all channels. In this case \(a = b = c\) and \(o_1 = o_2 = o_3 = 0\);

2. light intensity shift: there is an equal shift in image intensity values in all channels. In this case \(a = b = c = 1\) and \(o_1 = o_2 = o_3\);
3. light intensity change and shift: image values change by a constant factor in all channels and an equal shift in image intensity values is present in all channels. In this case $a = b = c$ and $o_1 = o_2 = o_3$.

4. light color change: image values change by different factors in each channel. In this case $o_1 = o_2 = o_3 = 0$.

5. light color change and shift: it corresponds to the full diagonal model.

Different color descriptors differ in their response to the illumination changes described above. A study of this behavior is presented in\(^{28}\). It is then extremely important to be able to classify illumination changes in the considered camera network. In the proposed approach the change in the color of clothes worn by the targets while they move from a camera view to another is analyzed. In order to do so, for each target two single-frame observations are extracted, one from camera $C_i$ and one from camera $C_j$, and the dominant color of the shirt observed by one camera is mapped with respect to the color observed by the other, and an evaluation of the linear curve that best fits the collected points is performed for each color channel.

Experimental tests performed on the available database has allowed to verify that for the recorded videos the R, G and B components almost overlap, that they are approximately shifted by the same value $(o_1 = o_2 = o_3 = 52)$ and that they are scaled by the same factor $(a = b = c = 0.72)$. Based on the data collected in the considered network it is possible to conclude that it is affected by light intensity change and shift.

The next step is based on the localization of shirt regions. To this aim, the proportion of the human body are studied. Anthropometry show that the average adult human figure is about 7 to 7.5 heads tall; starting from this consideration, the human body can be hypothetically divided in seven zones or Regions of Interest (ROI). The third zone is extended from the heart to the navel and for this reason, we will consider this part as the zone containing the shirt.

This method is prone to errors such as a wrong individuation of the ROI, detection of extraneous objects in the ROI, presence of multi-color clothes and unpredictable illumination changes, but it is reliable enough for understanding the nature of illumination changes.

6. CSD COLOR DESCRIPTOR FOR TARGET COLORS STRUCTURE

As mentioned before, the considered network is affected by light intensity change and shift, in order to re-identify objects through different cameras it is fundamental to find suitable descriptors that are invariant with respect to this kind of illumination change. To avoid ambiguity and errors in the re-identification system, caused by instability across the achromatic components and noise, HSV color space has been adopted, as suggested in\(^{28}\).

In this section we present a novel color descriptor, the CSD (Color Structure Descriptor), which is intensity scale and shift invariant.

This descriptor works in the RGB space therefore it is not affected by instability across the achromatic components. Thanks to this property this descriptor will be used in the following for reducing the number of possible matches.

The CSD allows to describe the difference between two regions of interest in term of dominant colors. In the considered case the shirt and the pants of the target are analyzed. As it has been explained in Section 5, in order to select the areas of interest the vertical proportions in the human body are used for partitioning the bounding box around each person in seven parts. Let us consider a target $\psi$ that crosses the camera $C_i$, the CSD of the generic frame $z$, is evaluated over $ROI^z_{i,\psi}$ and $ROI^b_{i,\psi}$, where the first region is represented by the foreground pixels that are inside the third part of the bounding box starting from the head and the second region is represented by the foreground pixels that are inside the fifth and the sixth regions of the bounding box starting from the head. Figure 2 shows the regions of interest on a target.

In order to simplify the notations, the subscript $z$ will be omitted in the following, and the performed considerations are applied to every frame.
Define as $DC^t_{i,\psi}$ and $DC^b_{i,\psi}$ the dominant colors evaluated on $ROI^t_{i,\psi}$ and $ROI^b_{i,\psi}$ respectively. In the RGB space:

$$DC^t_{i,\psi} = [R^t_{i,\psi}, G^t_{i,\psi}, B^t_{i,\psi}]$$
$$DC^b_{i,\psi} = [R^b_{i,\psi}, G^b_{i,\psi}, B^b_{i,\psi}].$$

Let $DDC_{i,\psi}$ be the difference between the dominant colors:

$$DDC_{i,\psi} = DC^t_{i,\psi} - DC^b_{i,\psi} = [R^{DD}_{i,\psi}, G^{DD}_{i,\psi}, B^{DD}_{i,\psi}]$$

where:

$$R^{DD}_{i,\psi} = R^t_{i,\psi} - R^b_{i,\psi}$$
$$G^{DD}_{i,\psi} = G^t_{i,\psi} - G^b_{i,\psi}$$
$$B^{DD}_{i,\psi} = B^t_{i,\psi} - B^b_{i,\psi}.$$ 

In this way the color structure is defined and the difference is invariant with respect to intensity shift. In order to guarantee also the invariance with respect to the intensity change we move to the RGB norm domain:

$$r_{CSD}, g_{CSD}, b_{CSD} = \frac{1}{R^{DD}_{i,\psi} + G^{DD}_{i,\psi} + B^{DD}_{i,\psi}} \begin{pmatrix} R^D_{i,\psi} \\ G^D_{i,\psi} \\ B^D_{i,\psi} \end{pmatrix}.$$ 

In this way the CSD for the target $\psi$ viewed from camera $C_i$ is given by:

$$CSD_{i,\psi} = [r_{CSD}, g_{CSD}, b_{CSD}]^{i,\psi}.$$ 

It is easy to demonstrate that this descriptor is invariant with respect to light illumination change and shift; for this reason we expect the distance between CSDs evaluated on the same target viewed from different cameras to be very low. It is worth remembering that the CSD describes a difference; for this reason it is likely to find a large number of couple of ROIs that are described by the same CSD values. The CSD is then used only to verify the coherence in the color structure and to delete those links in which this coherence is not present. The usage of CSD for a direct comparison would return a very noisy re-identification system.
7. TARGET COLOR STRUCTURE CONTROL

As mentioned before, a person re-identification problem can be seen as a graph problem; the observation of a target exiting from a camera is connected with all observations made from another camera, and to each edge in the graph is assigned a weight that represents the similarity score between connected observations. The CSD allows to reduce the number of links among observations.

Consider the link $k_{i,j}^{\psi,\varphi}$ between two observations $O_{i,\psi,z}$ and $O_{j,\beta,q}$. As previously explained, each observation consists of several single-frame observations; during the TCS (Target Color Structure) pruning step the distances in term of CSD between each single frame observation belonging to $O_{i,\psi,z}$ and each single frame observation belonging to $O_{j,\beta,q}$ is computed:

$$D_{k_{i,j}^{\psi,\varphi},z,q}^{\text{Str}} = ||CSD_{i,\psi,z} - CSD_{j,\beta,q}||.$$  \hspace{1cm} (8)

where $z$ and $q$ are two generic frames. The CSD between the two targets is chosen as:

$$D_{k_{i,j}^{\psi,\varphi}}^{\text{Str}} = \min(D_{k_{i,j}^{\psi,\varphi},z,q}^{\text{Str}}).$$  \hspace{1cm} (9)

Considering the invariant properties of the CSD, if $k_{i,j}^{\psi,\varphi}$ is a valid link, $D_{k_{i,j}^{\psi,\varphi}}^{\text{Str}} \approx 0$. The minimum value is selected in order to avoid errors due to the presence of extraneous objects, naked arms, occlusion or multicolor clothes.

The TCS control is performed by imposing that the link between two observations is kept if there is a coherence in the color structure worn by targets, i.e. if $D_{k_{i,j}^{\psi,\varphi}}^{\text{Str}}$ is minor than a threshold.

Figure 3 shows the set of distances $D^{\text{Str}} = \{D_{k_{i,j}^{\psi,\varphi},z,q}^{\text{Str}} | w = 1 : V_{i}, y = 1 : V_{j}\}$. i.e. the distances between the CSD value for each link that connects each observation taken from camera $C_{i}$ with each observation taken from camera $C_{j}$. The green dots represent the distances between consecutive observations of the same target, the red dots are the other ones. Figure 4 shows the FAR (False Acceptance Rate) and the FRR (False Rejection Rate) curves. The threshold has been set up at $T_{\text{TCS}} = 0.27$; for this value $FRR = 0$ and $FAR = 61.2$; with the implementation of the TCS control more than 38% of non-valid links are deleted. In order to evaluate these curves, a training phase is used in which 31% of the dataset is considered (15 targets).
8. RESULTS

In this section the evaluation of proposed approach is reported.

Test videos have been selected from two of the 40 stationary low-resolution cameras covering wide area of the UCSB campus. The dataset consists of 129 total observations over 772 frames: 56 observations are extracted from the data acquired by camera 1 while 73 observations from camera 2. In the performance evaluation, the 48 targets identified from camera 1 have been used as query objects. In the classical approach, that is without exploiting the space-time information, each query object has to be compared with the 73 objects tracked in camera 2.

As stated in the previous Sections, the basic idea of the proposed system is to increase the performance of a generic target re-identification system by considering as belonging to the same objects only those links presenting coherence in the color structure. This selection technique can be applied to any appearance-based approach. It is useful to note that, since the low resolution camera result in low definition of acquired faces and/or edges, the usage of face recognition based approaches is quite limited.

For this reason, in the following we show how the usage of TCS control can outperform the performances of the chosen appearance-based method. In our test, the 16 bin hue histogram comparison evaluated by the histogram intersection approach, has been selected as basic method.

In Figure 5, the recognition rate by the Cumulative Matching Characteristic (CMC) curve is shown. The rank score represents the number of matches considered. The CMC curve represents the expectation of finding the correct match in the top k matches.

As can be noticed, the usage of the TCS control allows to improve the performances of the basic method: considering the top rank, the recognition percentage is increased up to 4%. The color histogram matching rate, for top rank, increases from 22% up to 27%. The overall performance is increased up to 16% considering the first 11 candidates.

9. CONCLUSION

In this paper a methodology for improving the performances of appearance-based methods for person re-identification through a camera network, has been presented. The proposed approach is based on the usage of TCS control on the links: the link between two observations is considered belonging to the same object if there
is a coherence in the targets color structure. The color structure of a target is evaluated using a novel color descriptor called CSD (Color Structure Descriptor) which is intensity scale and shift invariant. The results show that this approach outperform the basic one up to 4% for the top rank and up to 16% considering the first 11 candidates. It is worth to consider that all the tests have been performed by considering low-resolution cameras, thus preventing target identification based on face recognition.

REFERENCES


