MVA2009 IAPR Conference on Machine Vision Applications, May 20-22, 2009, Yokohama, JAPAN

People Re-identification by Means of a Camera Network Using a Graph-based Approach

D-N. Truong Cong[1], L. Khoudour[1], C. Achard[2], P. Phothisane[2]
[1]French National Institute for Transport and Safety Research (INRETS)
[2]UPMC Univ Paris 06, Institute of Intelligent Systems and Robotics (ISIR)
{truong, louahdi.khoudour}@inrets.fr, catherine.achard@upmc.fr

Abstract

The problem described in this paper consists in re-identifying moving people in different sites with non-overlapping views. Our proposed framework relies on a spectral analysis of the appearance-based signatures extracted from a detected person in each sequence. First, we propose a new feature called "color-position" histogram combined with several illuminant invariant methods in order to characterize extracted silhouettes in static images. Then, a graph-based approach that characterizes the most useful information of video sequences is presented to compare sequences, two by two, and to make the final decision of re-identification. The global system is tested on a real and difficult data set composed of 40 individuals recorded in very different environments: indoors near windows and outdoors with very different lighting conditions. The experimental results show that our approach provides reliable results and is promising with regard to surveillance security applications.

1 Introduction

Over recent years, visual surveillance has gained more interest due to its important role in security. In an attempt to prevent the threats to security, such as aggressions against people, vandalism against property or acts of terrorism, more and more cameras are introduced to monitor public places. Manual supervision is a cumbersome task due to the large volumes of information and its expense for the hiring of staff. Therefore, it would be advantageous to automate this procedure by using a computer vision system that is able to extract the useful information represented in the video and to perform a certain task depending on the security scenario. One of the important tasks is to establish correspondence between observations of people who might appear and reappear at different times and across different cameras. In most cases, such a system relies on an appearance-based model that depends on several factors, such as illumination conditions, different camera angles and pose changes.

A significant amount of research has been carried out in the field of appearance-based person recognition. Nakajima et al. [1] presented a system which can recognize full-body people in indoor environments by using multi-class SVMs that were trained on color-based and shaped-based features extracted from the silhouette. Gheissari et al. [2] proposed a temporal signature which is invariant to the position of the body and the dynamic appearance of clothing within a video shot. Yang et al. [3] proposed an appearance model constructed by kernel density estimation. A key-frame selection and matching technique was presented in order to represent the information contained in video sequences and then to compare them.

The research presented in this paper is the development of a system that is able to re-identify moving people in a given site while observing them by means of a multi-camera system with different fields of view. First, we propose a new color-based signature in order to characterize the silhouettes in static images. Then, a graph-based approach is introduced to exploit the most useful information of video sequences and to improve the comparison of two sequences.

The organization of the article is as follows. In Section 2, our proposed signature and several illuminant normalizations are presented. In Section 3, after a few theoretical reminders on graph theory, we explain how to adapt the latter to our problematic. Section 4 presents global results on the performance of our system on a database of given facts. Finally, in Section 5, the conclusion and short-term perspectives are given.

2 Appearance-based feature extraction

The first step in our system consists in extracting from each frame a robust signature characterizing the passage of a person. To do this, a detection of moving areas, by using a background subtraction algorithm [4], combined with morphological operators is first carried out. Let us assume now that each person's silhouette is located in all the frames of a video sequence. Since the appearance of people is dominated by their clothes, color features are suitable for their description.

The most widely-used feature for describing the color of objects is color histograms that are resistant to deformable shapes and invariant to scale by normalization. The main drawback of color histograms is the lack of spatial information. This leads to the fact that they cannot discriminate between appearances that are the same in color distribution, but different in color structure. Several approaches have been proposed to include spatial information in the histogram format, such as multi-resolution histograms [5], spatiograms [6], and color/path-length feature [3].

In our research, we propose a new descriptor for static images called the "color-position" histogram. The idea of this signature is that the interested region (the silhouette in our case) can be horizontally decomposed in areas with homogeneous color. Thus, for estimating this new descriptor, the silhouette is first vertically divided in n equal parts. Then, the mean color is computed to characterize each part. The "color-position" histogram X is now composed of $n \times k$ values $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where n is the total number of equal parts of the silhouette and $\mathbf{x}_i$ is a vector of k color components. The advantages of such a signature are its consideration of the spatial information, its simple estimation and comparison, and its low memory consumption.

Since the color acquired by cameras is heavily dependent on several factors, such as the surface reflectance, illuminant color, response of the sensor, a color normalization procedure has to be carried out in order to obtain invariant signatures. Many methods have been proposed in the literature. In this paper, we only present the three invariances that lead to better results:

- Greyworld normalization [7] derived from the RGB space by dividing for each channel, the pixel value by the average of the image:

$$I^*_{R,G,B} = \frac{I_{R,G,B}}{mean\left(I_{R,G,B}\right)} \tag{1}$$

- Normalization using histogram equalization [8] is based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging illuminations. The *rank measure* for the level i is obtained with:

$$M_{R,G,B}(i) = \sum_{u=0}^{i} H_{R,G,B}(u) \Bigg/ \sum_{u=0}^{Nb} H_{R,G,B}(u) \tag{2}$$

where Nb is the number of quantization steps.

- Affine normalization is defined by:

$$I^*_{R,G,B} = \frac{I_{R,G,B} - \text{mean}(I_{R,G,B})}{\text{std}\left(I_{R,G,B}\right)} \tag{3}$$

These color normalizations are applied inside the silhouette of each person before computing its color-based signature. Thus, the output of this first step is the signatures that are invariant to lighting conditions and estimated on each frame. However, signature extracted from just one frame is not robust enough for comparing two video sequences, since they cannot contain all the appearance information in the sequence. A better solution is needed to characterize the useful information of a video sequence by using more than one frame. In the following section, we propose a graph-based approach that represents a set of signatures of two sequences in an embedded non-linear manifold without losing the original information. This procedure helps us to evaluate the similarity/dissimilarity between two video sequences by determining how separate two video sequences (two clusters) are in the graph and to make the final decision of re-identification.

3 Graph-based approach for sequence matching

3.1 Graph notion and random walks view

In this paper, we focus on the theory of random walks on graphs that can be used to describe the cohesion of a set of data points (a set of signatures, in our case). This technique preserves the local proximity among data points by first constructing a graph representation for the underlying manifold with vertices and edges. The vertices represent the data points, and the edges connecting the vertices represent the similarities between adjacent nodes. A random walk on the graph is a stochastic process which randomly jumps from vertex to vertex.

Given a set of signatures $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ extracted from m frames $\{I_1, I_2, \ldots, I_m\}$ belonging to two sequences, this set is associated to a complete neighborhood graph $G = (V, E)$ where each data point X_i (as well as each frame I_i) corresponds to a vertex v_i in this graph. Two vertices corresponding to two data points X_i and X_j are connected by an edge that is weighted by the similarity S_{ij} between two data points. Similarity S_{ij} is given by a Gaussian kernel $S_{ij} = exp\left(-\frac{d(X_i,X_j)^2}{\sigma^2}\right)$. Here, $d\left(X_i, X_j\right)$ is the distance between two signatures and the parameter σ is chosen as $\sigma = mean\left[d\left(X_i, X_j\right)\right], \forall i, j = 1, \ldots, m \ (i \neq j)$. Matrix $\mathbf{S} = \left[S_{ij}\right]_{i,j=1,\ldots,m}$ is called the *similarity matrix*. Let $\mathbf{D}$ denote the diagonal matrix with elements $D_{ii} = \sum_j S_{ij}$ where D_{ii} is the degree of a vertex $v_i \in V$. The transition probability of jumping in one iteration from vertex i to vertex j is given by $P_{ij} = S_{ij}/D_{ii}$. The transition matrix $\mathbf{P} = \left[P_{ij}\right]_{i,j=1,\ldots,m}$ of the random walk is thus defined by:

$$\mathbf{P} = \mathbf{D^{-1}S} \tag{4}$$

The set of the eigenvalues of $\mathbf{P}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, is usually called the *spectrum* of $\mathbf{P}$ (or the spectrum of the associated graph G). Since $\mathbf{P}$ is a right stochastic matrix and $P_{ij} > 0, \forall i, j = 1, ..., m$, the first eigenvalue λ_1 is 1 with the corresponding eigenvector $\gamma_1 = [1, 1, ..., 1]^T \Big/ \sqrt{m}$. In the application of spectral analysis for dimensionality reduction, the q first eigenvectors $\{\gamma_1, \gamma_2, \ldots, \gamma_q\}$ are used to create the new coordinate system for the set of data points. We can define a dimensionality reduction operator $h : X_i \to u_i = [\gamma_1(i), \ldots, \gamma_q(i)]$ where $\gamma_k(i)$ is the i^{th} coordinate of eigenvector γ_k.

Coming back to our problem, the set of signatures belonging to two video sequences is characterized by transition matrix $\mathbf{P}$ of the random walks on graph G. Note that the data set is composed of two known clusters, one for each video sequence. Our problem consists in evaluating how separate these two clusters are; or, in other words, in measuring the similarity between two video sequences. In the following, we present the solution of this problem based on the matrix perturbation theory and the relation between eigenvalues and eigenvectors of transition matrix $\mathbf{P}$.

3.2 Spectral analysis for sequence similarity measure

Let us consider the "ideal" case first, in which the data points within a cluster are infinitely far apart from all points of the second cluster. Assume also that data points $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ are ordered according to the cluster they belong to (i.e. the first points belong to video sequence a and the others to sequence b). Since two clusters are infinitively apart, the similarity matrix is a block diagonal matrix $\hat{\mathbf{S}} = \begin{bmatrix} \mathbf{S}^{(a)} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{(b)} \end{bmatrix}$ where $\mathbf{S}^{(a)}$ and $\mathbf{S}^{(b)}$ are the matrices of intra-cluster similarities of clusters a and b respectively.

Transition matrix $\hat{\mathbf{P}}$ is also a block-diagonal matrix. Its eigenvalues and eigenvectors are the union of the eigenvalues and eigenvectors of its blocks. Thus, the first two eigenvalues of $\hat{\mathbf{P}}$ are 1 and the matrix containing the first two corresponding eigenvectors as columns is $\hat{\Upsilon} = \begin{bmatrix} \gamma_1^{(a)} & \vec{0} \\ \vec{0} & \gamma_1^{(b)} \end{bmatrix}$. Note that $\gamma_1^{(a)}$ and $\gamma_1^{(b)}$ are the constant vectors. Points $\hat{u}_i$, which are defined as the i-th row of $\hat{\Upsilon}$, are identical for all data points X_i belonging to the same cluster ($\hat{u}_i = \begin{bmatrix} c^{(a)} & 0 \end{bmatrix}$ for cluster a and $\hat{u}_i = \begin{bmatrix} 0 & c^{(b)} \end{bmatrix}$ for cluster b). Therefore, the classification of points $\hat{u}_i$ leads to the clusters corresponding to the true clusters of the original data.

In general, the off-diagonal blocks of $\mathbf{S}$ and $\mathbf{P}$ are non-zero, i.e. the inter-cluster similarities are not exactly 0. The difference $\mathbf{E} = \mathbf{P} - \hat{\mathbf{P}}$ is considered as a perturbation. Matrix perturbation theory [9] demonstrates that the stability of the eigenvectors of a matrix is determined by the *eigengap*. More formally, the first k eigenvectors of $\hat{\mathbf{P}}$ will be stable to the perturbations of $\hat{\mathbf{P}}$ if and only if eigengap $\xi_k = |\lambda_k - \lambda_{k+1}|$ is large.

Let us apply this theorem to our problem. Transition matrix $\mathbf{P}$ is generally composed of non-zero off-diagonal blocks that are considered as perturbations. The more similar the two sequences, the larger the perturbations. Points X_i belonging to these two similar sequences are not well separated. This leads to the fact that points u_i are not well separated either. Thus, the first two eigenvectors of $\mathbf{P}$ are no longer close to the ideal eigenvectors. We assert that the first two eigenvectors are unstable to the changes of $\mathbf{P}$, and eigengap $\xi_2 = |\lambda_2 - \lambda_3|$ in this case is small.

If two sequences are quite different, we have a nearly ideal transition matrix $\mathbf{P}$ whose off-diagonal blocks are approximately 0. Points u_i might coincide with the ideal ones with a slight margin of error. In this case, the eigenvectors are stable and the eigengap is large.

Figure 1 illustrates the variation of eigengap according to the similarity of two sequences by representing the first five eigenvalues (left-hand diagrams) and the 2D visualizations in the coordinate $[\gamma_2, \gamma_1]$ (middle diagram) obtained by analyzing a set of signatures extracted from two sequences. The first data set consists of two sequences (10 frames per sequence) of the same person captured in different locations (first row). The second and third data sets are composed respectively of two sequences belonging to two people similarly dressed (second row) and two very different individuals (third row). We can remark that the eigengap in the first case is the smallest. The corresponding 2D visualization, on which the frames are represented by star points for one sequence and circle points for the other, shows the overlapping between the two sequences. In the second case, the two clusters are more separated. The eigengap of this case is larger than the first one, but it is still small due to the similarity between the clothing of the people. In the third case, the two clusters (sequences) are very distinct, and points u_i are almost fixed for each cluster. The eigengap is large enough to ensure the stability of the eigenvectors.

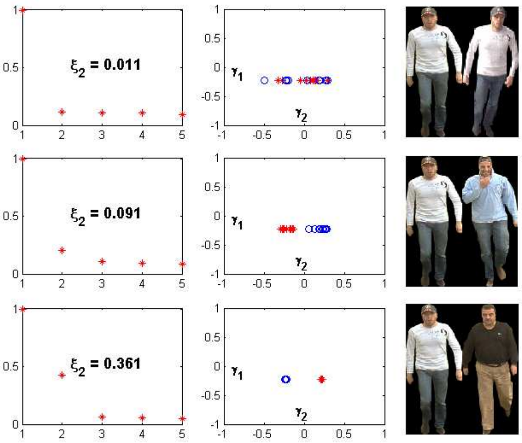

Figure 1: Illustration of the variation of eigengap according to the similarity of two sequences: the first five eigenvalues (left-hand diagrams), 2D visualizations with coordinates $[\gamma_2, \gamma_1]$ (middle diagrams) and representations of two sequences (right-hand figures).

Therefore, by constructing a random walk on the graph and computing spectrum of the associated graph, we can measure the similarity between two video sequences as the second eigengap $\xi_2 = |\lambda_2 - \lambda_3|$. Such a distance helps us to compare two video sequences and make the final decision of re-identification.

4 Experimental results

Our algorithms were entirely evaluated on real data sets containing video sequences of 40 people acquired in two environments. Each sequence is labeled as P-L where P encodes the person identification and L encodes the location. We have chosen very different locations (indoors in a hall near windows (L1) and outdoors (L2) with different lighting conditions). Figure 2 illustrates one of the forty people in these two different environments. In this figure, we notice that the color appearance is very different according to the location of the person.

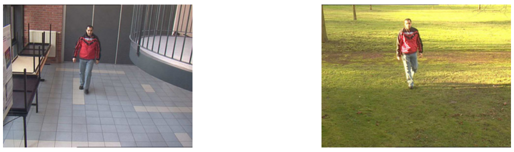

Figure 2: Illustrations of the large real database representing the same person in two different environments: indoors in a hall (left) and outdoors (right).

For each sequence, we extract ten frames regularly spaced in which people are viewed entirely. The objective of our system is to re-identify a person who has appeared in one surveillance environment among 40 sequences acquired in the other environment. Thus, sequence P[i]-L is compared with each sequence P[j]-L' (j=1,. . . ,40 & L' $\neq$ L). Distances obtained by these comparisons are classified in increasing order and the probability of correct re-identification at the top rank is calculated.

Table 1 reports the performance of our proposed approach. We note that there are two different situations in our experimentations. The first one is the re-identification of people who are captured in L1 and have already appeared in L2. In this case, each sequence captured in L1 is a query sequence and all the sequences in L2 are considered as the candidate set. The second situation is the opposite: each sequence captured in L2 is a query sequence and all the sequences in L1 are considered as the candidate set.

Table 1: Re-identification rate corresponding to four color spaces obtained by using the color-position histogram coupled with the graph-based approach.

	L1 matches with L2	L2 matches with L1
RGB space	70	72.5
Greyworld	95	100
Histogram equalization	97.5	97.5
Affine normalization	95	97.5

Thus, in table 1, the second column represents the rate of re-identification of the first situation and the third column represents the rate for the second situation. We notice that the rates are very satisfying: the best re-identification rate for the first situation is 97.5% and for the second situation is 100%. The invariant normalizations have actually improved the results in comparison to the RGB space.

5 Conclusion and perspectives

In this paper, we have presented a system that is able to track moving people in different sites while observing them through multiple cameras. We first propose a new descriptor for static images called the "color-position" histogram coupled with several illuminant normalizations. In order to further improve the appearance-based model of an individual, many images of a video sequence should be exploited. Hence, an algorithm which is based on the random walk on the graph is applied to compare two sequences and make the final decision of re-identification.

The global system was tested on a real and difficult data set composed of 40 individuals filmed at two different locations: indoors near windows and outdoors with very different lighting conditions. The experimental results have shown that our proposed approach provides reliable results. These results, which are the fruit of the combination of the color-position signature, the graph-based procedure and the illuminant invariance, are very satisfying.

Our framework needs to be evaluated more intensively. A good occasion will be to test it on people tracking in the transport environment in the framework of an European project. One of the tasks within this project is the study and development of automatic surveillance functions adapted to the context of the inside of an operating train. On-board automatic video surveillance is a challenge due to the difficulties in dealing with fast illumination variations, reflections, vibrations, high people density and static/dynamic occlusions that perturb actual video interpretation tools.

References

[1] C. Nakajima, M. Pontil, M. Heisele, and T. Poggio. Full body person recognition system. *Pattern Recognition*, 36(9):1997–2006, 2003.

[2] N. Gheissari, T.B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, Washington, DC, USA, 2006. IEEE Computer Society.

[3] Y. Yu, D. Harwood, K. Yoon, and L.S. Davis. Human appearance modeling for matching across video sequences. *Machine Vision and Applications*, 18(3):139–149, 2007.

[4] K. Kim, TH Chalidabhongse, D. Harwood, and L. Davis. Background modeling and subtraction by codebook construction. In *International Conference on Image Processing, ICIP'04.*, volume 5, 2004.

[5] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Spatial information in multiresolution histograms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[6] S.T. Birchfield and S. Rangarajan. Spatiograms versus histograms for region-based tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1158–1163, 2005.

[7] G. Buchsbaum. A spatial processor model for object color perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980.

[8] G.D. Finlayson, S. Hordley, G. Schaefer, and G. Yun Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38(2):179–190, 2005.

[9] G.W. Stewart and J. Sun. *Matrix perturbation theory.* Academic Press, 1990.