Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

People re-identification by spectral classification of silhouettes

D.-N. Truong Cong [a,*], L. Khoudour [a], C. Achard [b], C. Meurie [c], O. Lezoray [d]

[a] The French National Institute for Transport and Safety Research (INRETS), LEOST, 20, rue Elisée Reclus, 59650 Villeneuve d'Ascq Cedex, France
[b] UMPC Univ Paris 06, ISIR, UMR 7222, France
[c] University of Technology of Belfort Montbéliard, SeT, France
[d] University of Caen Basse-Normandie, GREYC UMR CNRS 6072, France

ARTICLE INFO

Article history:
Received 30 October 2008
Received in revised form
31 May 2009
Accepted 4 September 2009

Keywords:
Surveillance systems
Person re-identification
People tracking
Spectral analysis
Support vector machines
Color invariant

ABSTRACT

The problem described in this paper consists in re-identifying moving people in different sites which are completely covered with non-overlapping cameras. Our proposed framework relies on the spectral classification of the appearance-based signatures extracted from the detected person in each sequence. We first propose a new feature called "color-position" histogram combined with several illumination invariant methods in order to characterize the silhouettes in static images. Then, we develop an algorithm based on spectral analysis and support vector machines (SVM) for the re-identification of people. The performance of our system is evaluated on real datasets collected on INRETS premises. The experimental results show that our approach provides promising results for security applications.

1. Introduction

Nowadays, there is no doubt that security should be a major worry for the actors of public transport (travelers, staff, operating companies, governments). Each network or country has established measures according to their knowledge of these problems, to local conditions, and cultural traditions (for example: attitudes and legal limits relative to private life). Timely detection and intervention are needed in the case of threats for security, such as aggressions against people, vandalism against property, acts of terrorism, accidents and major catastrophes such as fires. The Closed-Circuit TeleVision (CCTV) coverage, which is considered as an essential element by several networks of large and middle-size cities, local authorities and police forces, has improved unceasingly. For instance, it was estimated that more than a million cameras are in public places in the United Kingdom and that on average, an individual is "seen" by 300 cameras in only one day in London.

However, the lack of staff limits drastically the general use of CCTV, especially if these systems must be used for prevention, rather than to react after the detection of accidents. It is usual that a human operator, responsible for a video surveillance system, should have to manage simultaneously 20–40 video sources. It brings new difficulties in defining the suitable procedures capable of managing the large volumes of information produced by such systems. When raw video data are available, one must automatically identify incidents as well as dangerous and potentially dangerous situations. Indeed, it is essential to avoid the visual excess to which human operators are currently exposed.

The research presented in this paper is within the framework of BOSS European project [1] (on BOard wireless Secured video Surveillance) which aims at developing a multi-camera vision system specified to monitor, detect and recognize abnormal events occurring on-board trains. One of the important tasks of such a system is to establish correspondence between observations of people

* Corresponding author. Tel.: +33 3 20 43 83 43; fax: +33 3 20 43 83 59.
E-mail address: truong@inrets.fr (D.-N. Truong Cong).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.09.005

over different camera views located at different physical sites. In most cases, such a task relies on the appearance-based models of moving people that may vary depending on several factors, such as illumination conditions, camera angles and pose changes.

In this paper, we propose a particular function between two cameras in order to re-identify a person who has appeared in the field of one camera and then reappears in front of another camera. Our proposed approach consists of several steps. First, we compute invariant features (also called signatures) in order to characterize the silhouettes in static images. Then, a graph-based approach is introduced to reduce the effective working space and realize the comparison of two video sequences (two passages). The performance of our system is evaluated on a real dataset containing 40 people filmed in two different environments (one indoors and one outdoors).

One of the originalities of our research is the tracking of people that represent in the image processing field, what are called deformable shapes. The second originality is the developed algorithms based on spectral analysis and support vector machines (SVM) for the re-identification of people as they move from one location to another. Lastly, the third strong point is that the algorithm is fully illuminant invariant.

The organization of the article is as follows: after this introduction, we will find in Section 2 a short state of the art on video sequence comparison. Section 3 describes how the invariant signature of a detected person is generated. In Section 4, after a few theoretical reminders on spectral analysis, we explain how we adapt the latter to our problematic. The first illustrated results allow us to establish a good discrimination between individuals. In Section 5, we briefly describe the main concepts of SVM and their application to our problem. In fact, the use of SVM is an interested step that complements spectral analysis to perform re-identification. Section 6 presents global results on the performance of our system on a real dataset. Finally, in Section 7, conclusions and important short-term perspectives are given.

2. State of the art on video sequence comparison

Over the past several years, a significant amount of research has been carried out in the field of object recognition by comparing video sequences. It is usual to describe the color-based features of video sequences using a set of key frames that describes well an entire video sequence. Several techniques of key frame selection from video sequences have been proposed so far. Ueda et al. [2] used the first and last frames of each sequence as two key frames. Ferman et al. [3] clustered the frames in each sequence. The closest frame to the center of the largest cluster is selected as the key frame for that shot. Sun et al. [4] divided a video sequence into intervals which are determined by computing the largest dissimilarity between the first and last frames. Girgensohn et al. [5] determined the key frames by clustering the frames in a video shot and by selecting the most representative frame for each cluster. Yang et al. [6] proposed a key frame selection process based on a comparison of the distances of the current key frame to the following frames with a given threshold. Although the latter key frame selection techniques are computationally inexpensive, the video sequence description they provide varies significantly with the selection criterion.

Given the drawback of key frame extraction methods, a preferable approach is to consider the characteristics of all the frames within a sequence and to compute a single compound signature of the sequence. Ferman et al. [7,8] proposed various histogram-based color descriptors to represent the color properties of a sequence. Leclercq et al. [9] proposed to use co-occurrence matrices to have a spatial distribution of the pixels in a shape. They then used principal component analysis (PCA) for dimensionality reduction and final classification. Gheissari et al. [10] proposed a temporal signature which is invariant to the position of the body and the dynamic appearance of clothing within a video shot.

3. Signature generation

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 background subtraction, combined with a shadow elimination algorithm is first carried out [11,12]. 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. Several tools can then be used, such as the *color histogram* [13] that is the most commonly used structure to represent global image features. It is invariant to translation, rotation and can become invariant to scale by normalization. The undeniable advantage of the *color path length feature* [14] is its ability to include some spatial information: each pixel inside the silhouette is represented by a feature vector (x, l), where x is the color value and l is the length between an anchor point (the top of the head) and the pixel. The distribution of $p(x, l)$ is then estimated with a 2D histogram. We can lastly cite *spatiograms* [15], which are a generalization of histograms including higher order spatial moments. For example, the second-order spatiogram contains, for each histogram bin, the spatial mean and covariance.

In our research, we propose a new descriptor for static images called the "color-position" histogram (Fig. 1). This is really easy to estimate because the silhouette is first vertically divided into n equal parts. Then, the mean color is computed to characterize each part. The "color-position" histogram is then composed of $n \times 3$ values (while working with three color channels). Compared to the classical color histogram, it leads to better results (thanks to the spatial information) and uses less memory. Its advantages regarding the color path length feature are a faster estimation and lower memory consumption. Furthermore, this new feature is more homogeneous than the spatiogram; this leads to simple and more reliable measures to compare two silhouettes.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

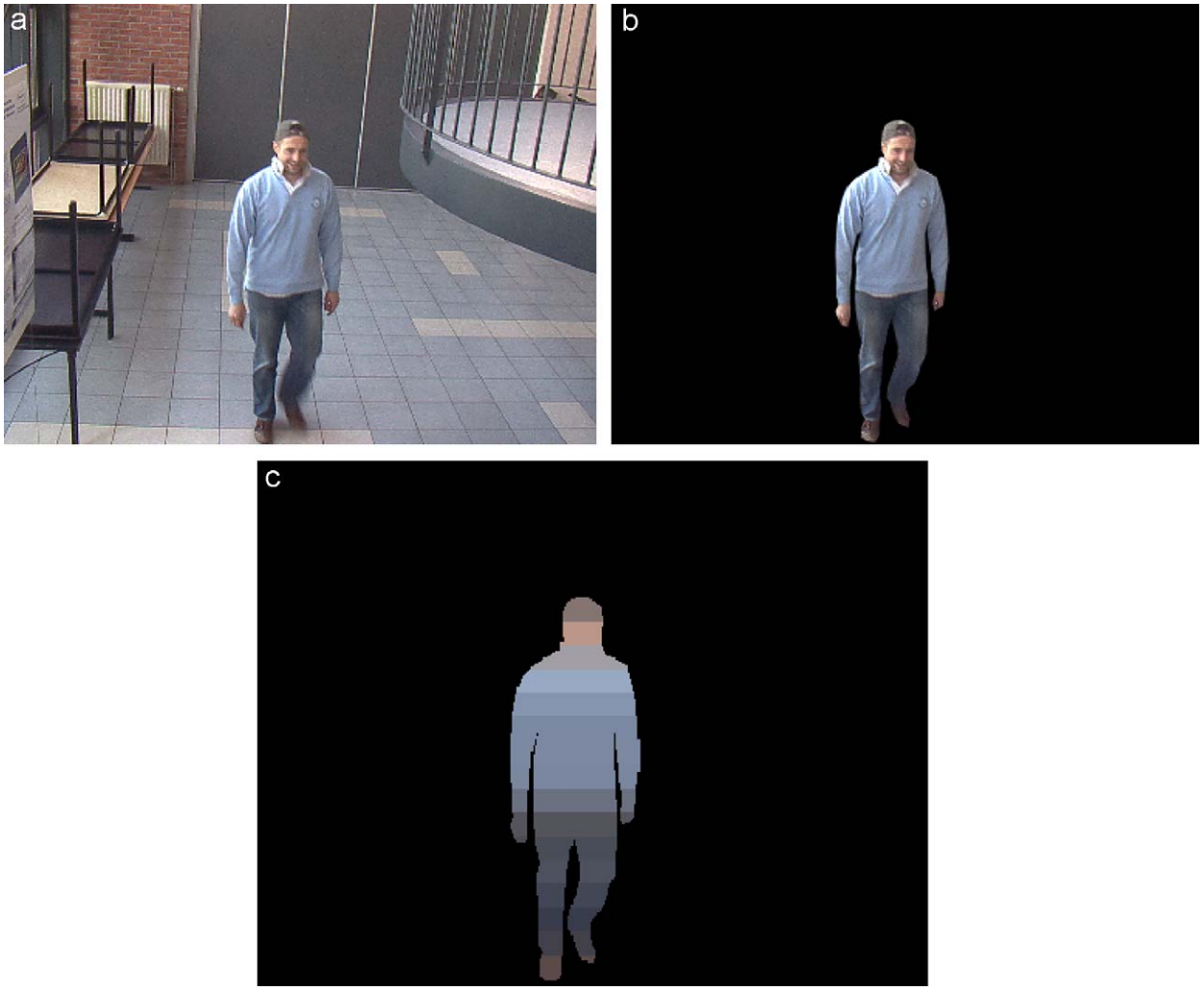

Fig. 1. Color-position histogram: original image (a), localization of the silhouette (b), color distribution in the silhouette (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Unfortunately, the color acquired by cameras is heavily dependent on several factors, such as the surface reflectance, illuminant color, lighting geometry, response of the sensor, etc. and preliminary processings have to be introduced to obtain invariant signatures.

Several normalizations have been proposed in literature and we tested most of them. We only cite the most interesting as the *chromaticity space* derived from the RGB space from:

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B} \tag{1}$$

It is very simple and is independent of illuminant intensity (but not illuminant color).

Greyworld normalization [16] consists in dividing for each channel, the pixel value by the average of the image (or in a given area):

$$R' = \frac{R}{\text{mean}(R)}, \quad G' = \frac{G}{\text{mean}(G)}, \quad B' = \frac{B}{\text{mean}(B)} \tag{2}$$

This normalization is derived from the diagonal model of color change proposed by Finlayson et al. [17]. Invariance is obtained according to illuminant color (but not to illuminant intensity).

To be invariant to both intensity and color changes, Finlayson et al. have introduced the *comprehensive normalization* procedure [17], which is an iterative algorithm with two steps in the principal loop, one for each invariance.

A new feature based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging illuminations has also been introduced [18]. The *rank measure* for the level i and the channel k is obtained with

$$M_k(i) = \sum_{u=0}^{i} H_k(u) \Big/ \sum_{u=0}^{i} H_k(u) \tag{3}$$

where Nb is the number of quantization steps and $H_k(\cdot)$ is the histogram for the channel k.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

Lastly, we have tested the *affine normalization* defined by

$$R' = \frac{R - \text{mean}(R)}{\text{std}(R)}, \quad G' = \frac{G - \text{mean}(G)}{\text{std}(G)}, \quad B' = \frac{R - \text{mean}(B)}{\text{std}(B)} \tag{4}$$

For all these methods, the color normalization is applied inside the silhouette of each person before computing its color-position histogram. A comparative study of the different normalization procedures will be presented in Section 6.

The output of this first step is a color-position histogram, invariant to lighting conditions and estimated on each frame. However, the signature extracted from just one frame is not robust enough for comparing two image sequences. A stronger solution is needed to characterize the whole sequence. In the following section, we will introduce a graph-based approach that can reduce the dimensionality of our dataset (set of signatures of a sequence) without losing useful information and obtain a single representation of a whole sequence.

4. Dimensionality reduction

4.1. Overview

High-dimensional data, meaning data that require several dimensions to represent, can be difficult to interpret and process. One approach to tackle this problem is to assume that the data of interest lies on an embedded non-linear manifold within the higher dimensional space. If the manifold is of low enough dimension then the data can be visualized in the low dimensional space. Spectral methods have recently emerged as a powerful tool for non-linear dimensionality reduction and manifold learning [19,20]. Each input example is then associated with a low-dimensional representation that corresponds to its estimated coordinates on the manifold. Dimensionality reduction can yield to a new representation that preserves almost all the original information while this new representation can also ease learning and improve generalization in a supervised learning process. In addition to being useful as a preprocessing step for supervised learning, non-linear dimensionality reduction is often used for data analysis and visualization, since visualizing the projections of the data can help to better understand it. In the last few years, many unsupervised learning algorithms have been proposed which share the use of an eigen-decomposition for obtaining a lower-dimensional embedding of the data that characterizes a non-linear manifold near which the data would lie: locally linear embedding (LLE) [21], Isomap [22], Laplacian eigenmaps [23], diffusion maps [24] and many variants of spectral analysis [25,26].

In this paper, we only focus on graph-based methods for non-linear dimensionality reduction. Sometimes called diffusion maps, Laplacian eigenmaps or spectral analysis, these manifold-learning techniques preserve the local proximity between 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. If properly normalized, these edge weights can be interpreted as transition probabilities for a random walk on the graph. After representing the graph with a matrix, the spectral properties of this matrix are used to embed the data points into a lower dimensional space, and gain insight into the geometry of the dataset.

Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \in \mathbb{R}^n$ be m sample vectors. Given a neighborhood graph G associated to these vectors, one considers its adjacency matrix W where weights W_{ij} are given by a Gaussian kernel $W_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$. Let D denote the diagonal matrix with elements $D_{ii} = \sum_j W_{ij}$ and Δ denote the un-normalized Laplacian defined by $\Delta = D - W$. The dimensionality reduction consists in searching for a new representation $\{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ with $\varphi_i \in \mathbb{R}^m$, obtained by minimizing

$$\frac{1}{2}\sum_{ij} \|\varphi_i - \varphi_j\|_2 W_{ij} = \text{Tr}(\mathbf{Y}^T \Delta \mathbf{Y}) \quad \text{with } \mathbf{Y} = [\varphi_1, \varphi_2, \ldots, \varphi_m] \tag{5}$$

This cost function encourages nearby sample vectors to be mapped to nearby outputs. This is achieved by finding the eigenvectors $\varphi_1, \varphi_2, \ldots, \varphi_m$ of matrix Δ. Dimensionality reduction is obtained by considering the q lowest eigenvectors (the first eigenvector being discarded) with $q \ll n$. Therefore, we can define a dimensionality reduction operator $h : \mathbf{x}_i \to (\varphi_2(i), \ldots, \varphi_q(i))$ where $\varphi_k(i)$ is the i th coordinate of eigenvector φ_k. When the graph G is a neighborhood graph (e.g. a k nearest neighbor graph), this dimensionality reduction is called Laplacian eigenmaps [23]. When the graph G is a complete graph, this dimensionality reduction is called diffusion maps [24]. Both methods are equivalent (up to some normalization) and very close to spectral analysis [20]. In the rest of the paper, we will use the term spectral analysis to denote a dimensionality reduction performed by the above-mentioned graph-based approach.

4.2. Silhouette categorization

In this section, we present our framework based on spectral analysis that is able to reduce the dimensionality of an image set and provides a new 2D visualization. This approach enables us to visualize the images of two sequences in a 2D space and thus helps us to interpret them more easily.

Given an image set S consisting of m images belonging to two sequences $S = \{\mathbf{I}_1, \mathbf{I}_2 \ldots, \mathbf{I}_m\}$, the first step of our framework is to extract the invariant signature "color-position" described in Section 3. This leads to a new set of vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \in R^k$, where each vector $\mathbf{x}_i$ corresponds to the image $\mathbf{I}_i$. We now associate to the set of vectors X a complete neighborhood graph $G = (V, E)$ where each vector $\mathbf{x}_i$ (as well as each image $\mathbf{I}_i$) corresponds to a vertex v_i in this graph. Two vertices corresponding to two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by an edge that is weighted by $W_{ij} = \exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/\sigma^2)$. Here, we use the L^1 norm for computing the distance between two characteristic vectors $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$. The

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

parameter σ is chosen as $\sigma = \text{mean}[d(\mathbf{x}_i, \mathbf{x}_j)]$, $\forall i,j = 1,\ldots,m$ $(i \neq j)$. Ideally, W_{ij} is large when images indexed by i and j are preferably in the same sequence, and is small otherwise. Now, we can compute the un-normalized Laplacian Δ and produce the eigenstructure of Δ. The eigenvectors $(\varphi_1, \varphi_2, \ldots, \varphi_m)$ provide a new coordinate for the image set.

Dimensionality reduction is obtained by considering the q lowest eigenvectors with $q \ll n$. Choosing number q could be a problem. A solution proved by von Luxburg et al. [27] is that the eigenvalues corresponding to the eigenvectors used for dimensionality reduction and then for spectral clustering must be significantly below the minimal degree in the graph (i.e. $\lambda_i \ll \min_{j=1,\ldots,n} D_{jj}$, $\forall i = 1,\ldots,q$). The other eigenvectors which correspond to eigenvalues with $\lambda \geq \min D_{jj}$ are almost Dirac functions. Within the framework of our approach, two eigenvectors (φ_2, φ_3) whose eigenvalues are significantly below $\min D_{jj}$ are used for creating a 2D projection suitable for the visualization of the whole image set. Each image $\mathbf{I}_i$ is now represented by point $\mathbf{u}_i = (\varphi_2(i), \varphi_3(i))$ in the 2D Euclidean space.

In order to illustrate the output of the spectral analysis and demonstrate that it is a good representation for visualizing and comparing two sequences, we carried out several tests. The first is achieved by applying the spectral analysis to an *image set* composed of two sequences (10 frames per sequence) representing two people differently dressed.

Fig. 2 presents the 2D space (φ_2, φ_3) in which the set of frames of two sequences is plotted. On the left-hand diagram, the frames of the two sequences are illustrated by star points (blue for one person and red for the other) while on the right-hand diagram the points are directly illustrated by the corresponding silhouettes. According to the results shown in Fig. 2, we notice that the image set of this experiment contains two well-separated clusters with a large gap between both (i.e. the space between the two clusters is large). In other words, as the image set is naturally partitioned into two disjoint classes in the (φ_2, φ_3) space, we can assert that the two tested sequences represent two different people. In this case, only φ_2 eigenvector is sufficient to perform the clustering. This corresponds to use the sign of φ_2 and is equivalent to use the normalized cut algorithm [26].

The second trial is carried out with sequences of two different people very similarly dressed (they both wear a white T-shirt and blue jeans). Note that these two sequences are captured by two cameras located at two different sites (indoors in a hall near windows and outdoors with natural light). The result is shown in Fig. 3. For this image set, we can see that the two clusters are now less easily identifiable because, even if the two sequences represent two different individuals, their colorimetric appearances are very similar. There is still a gap between the two clusters, but it is not so easy to split the 20 frames into two groups without any prior knowledge. It is worth to note that here the normalized cut criterion is not accurate enough to perform the clustering.

The last trial image set consists of two sequences of the same person captured in different locations: indoors, in a hall near windows and, outdoors, in a garden. Shooting environments are completely different in terms of lighting, background, and so on. The mapping in Fig. 4 shows that the clusters of the two image sequences strongly overlap. There are no longer clear clusters and well-defined gap in this image set. This means that, in spite of the different environments, the two groups of frames are recognized as similar and, in other words, correspond to the same person.

Thus, these first experimental results illustrate that, by using the invariant signature "color-position" histogram to create the set of characteristic vectors of two test sequences and by applying spectral analysis, we can obtain a new visualization of an image set that helps us to determine the gap (the distance) between two image

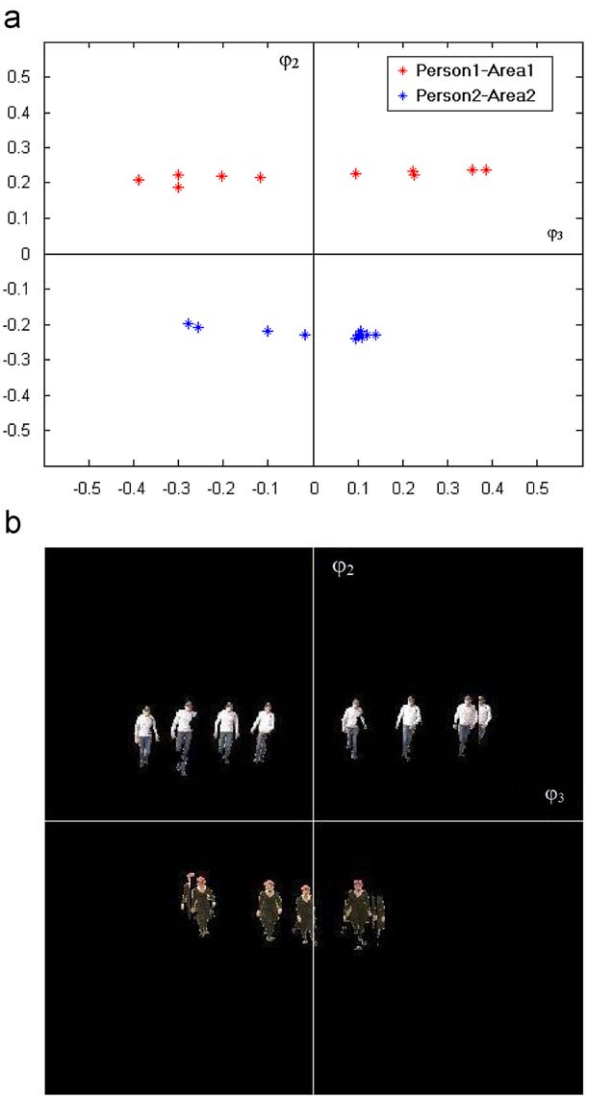

Fig. 2. Visualization of two sequences containing two people differently dressed in the 2D space represented by (φ_2, φ_3) (one wears a white T-shirt and blue jeans, the other wears a black dress). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

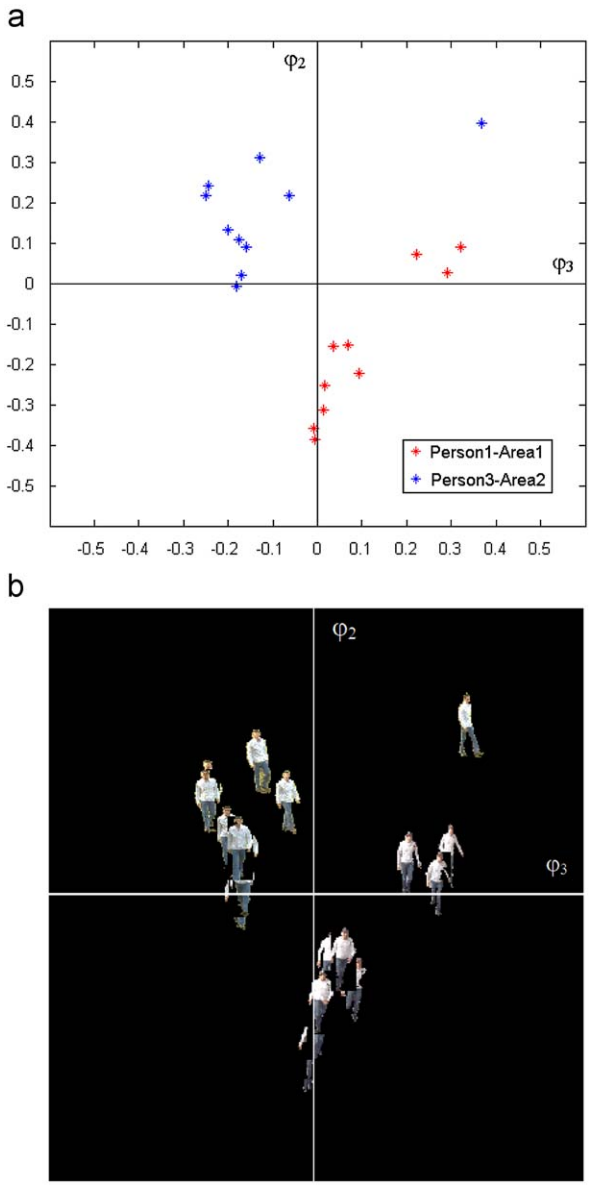

Fig. 3. Visualization of two sequences belonging to two different people similarly dressed in the 2D space represented by (φ_2, φ_3) (both wear a white T-shirt and blue jeans). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

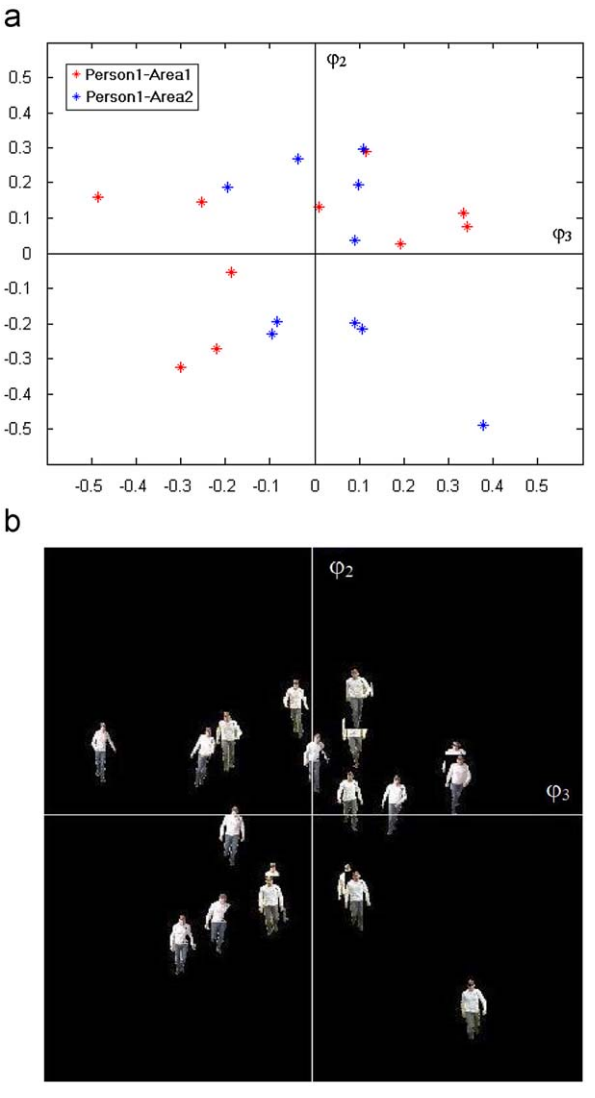

Fig. 4. Visualization of two sequences describing the same person in a different location in the 2D space represented by (φ_2, φ_3).

sequences. The more dissimilar the two sequences, the larger the gap. Because our objective is to re-identify people (recognize them from one camera to another) based on their color appearance, these results are very satisfactory and encourage us to continue in this way.

Spectral analysis is a very important step allowing a dimensional reduction without losing too much information included in the data. At this step, a higher level module has to be introduced to take a final decision of re-identification when comparing two sequences. Different parametric, non-parametric, and discriminating methods have been considered and the SVM seems to be appropriate for our problem. In the following section, we will briefly describe the main concepts of SVM and their specific application to our problem.

5. Application of SVM in measuring the similarity of two sequences

In Section 4, we described how an image set can be mapped into a 2D plane by using spectral dimensionality reduction. Several experimental results showed that the new coordinate system is a good representation for visualizing the image set. Moreover, it introduces a gap between two clusters that can be used to solve our objective of re-identification. We present in this section the application of SVM [28] (see Appendix for more details) to define the gap between two clusters (two

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

groups of frames, for instance), and to compute the distance between the two sequences.

Let $\mathbf{u}_i \in R^2$ be the vector obtained by spectral analysis corresponding to the i-th image of the tested image set. Each vector is labeled by a class $v_i \in \{-1, +1\}$ according to the sequences it belongs to. The linear SVM whose kernel function $K(\cdot,\cdot)$ is defined as $K(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{u}_i^T \mathbf{u}_j$ is now applied directly to the input–output set $(\mathbf{u}_i, v_i)$ in order to determine the optimal hyperplane $\mathbf{w} \cdot \phi(\mathbf{w}) + b = 0$ which separates the two classes with the widest margin. Computing this hyperplane is equivalent to minimize the following optimization problem [29]:

$$\mathcal{V}(\mathbf{w}, b, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{m} \xi_i\right) \tag{6}$$

where the constraint $\forall_{i=1}^{m} : v_i[\mathbf{w} \cdot \phi(\mathbf{u}_i) + b] \geq 1 - \xi_i, \xi_i \geq 0$ requires that all training examples are correctly classified up to some slack ξ and C is a parameter allowing trading-off between training errors and model complexity.

We first discuss the case where the training dataset is linearly separable (Figs. 2 and 3). This means that it is possible to find an optimal hyperplane which separates two classes without error (i.e. there is no slack ξ in classification). The distance between two image sequences in this case is defined as the optimal margin $2/\|\mathbf{w}\|$ obtained by the SVM. Fig. 5 shows the results obtained by applying SVM to the image sets shown in Figs. 2 and 3. We notice that the distance between the two image sequences in the first test (Fig. 5a) is larger than in the second (Fig. 5b). This means that the more different the appearances of two individuals are, the larger the distance between two image sequences.

The above discussion has been restricted to the case where the image set is linearly separable. For the non-separable image set, there are always several misclassification errors which are measured by the slack ξ_i. The result of the classification in this case depends on parameter C (see Eq. (6)) which corresponds to the degree of penalty assigned to an error. In our algorithm, we choose C equal to infinity. For such a value of C, the solution of SVM converges towards the solution obtained by the optimal separating hyperplane for the non-separable dataset.

Fig. 6 shows the result obtained by applying SVM to the image set shown in Fig. 4. The errors of classification are represented by a surrounding circle in the diagram. We notice that, for such a dataset, we cannot find two hyperplanes H_1 and H_2 which separate the two sequences according to the linear model. The distance between two image sequences in this case can be considered equal to 0, or, in other words, these two sequences represent the same person in our case. However, such an assessment can result in two possible cases: a true re-identification (true positive) if two image sequences are from the same person and a false re-identification (false positive or type 1 error) if two image sequences are from two different people.

In order to further describe the characteristic of two image sequences in the non-separable case, we introduce a notion of "mixture score" which is defined as

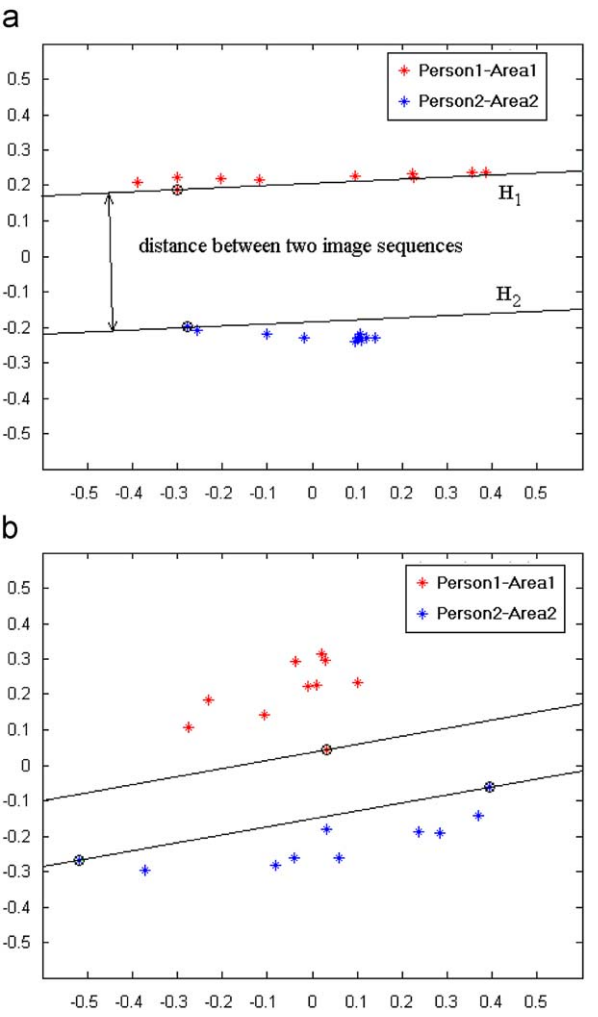

Fig. 5. Linear separating hyperplanes for the image set of Figs. 2 and 3.

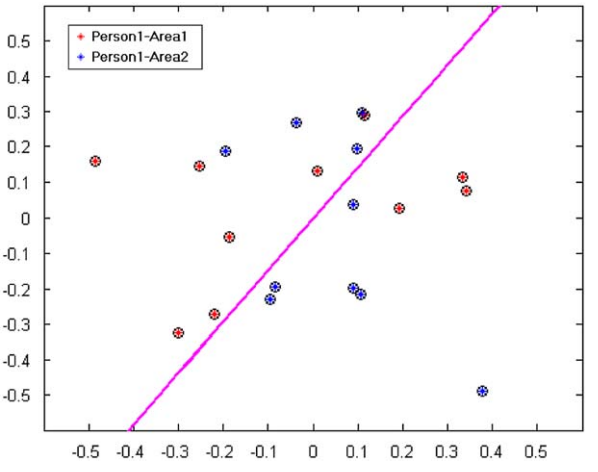

Fig. 6. Application of SVM for the image set of Fig. 4.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

$s = -\sum_{i=1}^{k} \xi_i / \|\mathbf{w}\|$. The more misclassification errors there are, the smaller the mixture score. This notion can be used as a complementary condition for comparing two sequences in case there are many pairs of sequences which cannot be separated by a linear model.

6. Experimental results

As mentioned above, our research aims to set up an on-board surveillance system that is able to re-identify a person through multiple cameras with different fields of vision. Before collecting a real on-board dataset, a large database containing video sequences of 40 people acquired in INRETS premises was collected for the evaluation of our algorithms. We have chosen two different locations (indoors in a hall near windows and outdoors with natural light) to set up these two cameras. Fig. 7 illustrates one of the 40 people in these two locations. We notice that the color appearance is very different according to the real scene illumination conditions.

For each video sequence, the silhouette of the moving person is extracted by using the background subtraction technique, combined with a shadow elimination algorithm [11,12] and morphological operators (erosion and dilation). A set of key frames in which people are entirely viewed is then extracted in order to characterize the passage of an individual. For this database, we have chosen to extract 10 key frames per passage and per location. Such a number of frames is sufficient for describing the characteristics of a passage in front of camera and for ensuring an adequate time processing. Fig. 8 illustrates the key frames extracted from two sequences of the same person.

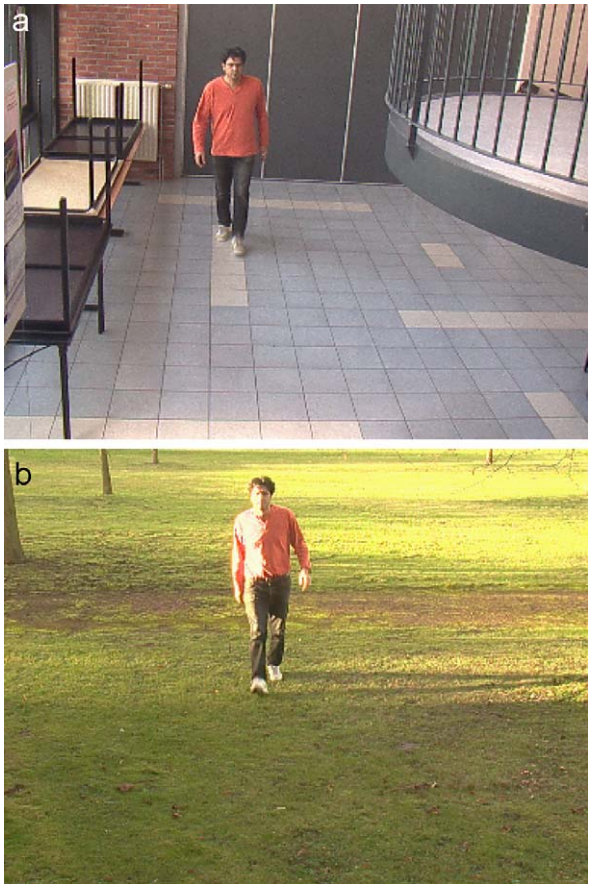

Fig. 7. Illustrations of the large real database representing the same person in two different environments: indoors in a hall (left) and outdoors (right).

For each query passage in front of one camera, the distances between the query passage and each of the candidate passages of the other camera are calculated by applying the spectral analysis for dimensional reduction and the SVM for similarity measure. A decision threshold is chosen; distances below the threshold indicate a re-identification (score = 1). This means that two passages belong to the same person. If the distance is above the threshold, this means it is a distinction and these two test passages belong to two different individuals (score = 0).

Since there are 40 video sequences for each location, 40×40 distances (i.e. dissimilarities between two video sequences) are calculated and then compared with the threshold. The resulting scores can be arranged in a 40×40 score matrix. An ideal score matrix is one whose diagonal elements are 1 (true re-identification) and whose off-diagonal elements are 0 (true distinction). In fact, a real re-identification system can give one of four possible results:

- True re-identification (also known as true match, true positive): the system declares a re-identification (score = 1) when the two passages belong to the same person (the diagonal).
- True distinction (also known as true non-match, true negative): the system declares a distinction (score = 0) when the two passages represent two different people (the off-diagonal).
- False re-identification (also known as false positive, false match or type II error): the system declares a re-identification (score = 1) when the two passages represent two different people (the off-diagonal).
- False distinction (also known as false negative, false non-match or type I error): the system declares a distinction (score = 0) when the two passages represent the same person (the diagonal).

Hence, these four possible rates (true re-identification rate, TRR, true distinction rate, TDR, false re-identification rate, FRR, and false distinction rate, FDR) can be calculated from the score matrix and are functions of the threshold which can be changed according to the context of utilization of the system. In our system, we choose the optimal threshold by referring to the equal true rate (ETR) point which assumes the equality of TRR and TDR. Two such rates can be calculated from the score matrix by using the following definitions:

$$\mathrm{TRR} = \frac{\sum_{k=1}^{N} (\mathrm{score}_{kk} = 1)}{N} \tag{7}$$

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

Fig. 8. Example of frame extractions for two sequences of the same person in two different locations (outdoors in a garden (first row); indoors in a hall (second row)).

$$\mathrm{TDR} = \frac{\sum_{k=1}^{N}\sum_{l=1}^{N}(\mathrm{score}_{kl} = 0, k \neq l)}{\mathrm{N}(N-1)} \tag{8}$$

where N is the number of people in the database ($N = 40$ in our case).

In Fig. 9, we find the TRR (in red) and TDR (in blue) according to given thresholds. For instance, in the first part which corresponds to RGB space, i.e. without invariant, the ideal setting of the thresholds leads to an 86% rate either for distinction or re-identification. This result represents the crossing of the two curves. This means that re-identification and distinction are of equal importance from the user's point of view.

Table 1 summarizes the comparative results obtained at the optimal points corresponding to RGB space and five illuminant normalizations. We can notice that, except for two invariants which are chromaticity space and comprehensive normalization, the others have actually improved the results in comparison to the RGB space. In particular, Greyworld illuminant invariant is the one which leads to the best performance (TRR increases from 86% to 95%).

Another way of showing more clearly the performance of the system combined with the invariants is to use a ROC (receiver operating characteristic) curve as illustrated in Fig. 10. In this figure, we can find a plot of TRR versus FRR as the value of threshold varies for the RGB space and the five invariants used. The closer the curve approaches the top left-hand corner of the plot, the better the method is. The ETR line is also represented in this figure in order to determine the six optimal points (the crossing between each curve and the ETR line). Based on the results presented in Fig. 10, we can confirm that the Greyworld normalization is the best method compared to the others. Its ROC curve is the closest curve to the top left-hand corner of the plot and its ETR point gives us a very satisfying rate of re-identification.

Here, we note that the TRRs of our system can be regulated by the decision threshold according to the context of utilization of the system. Another approach for solving our problem of re-identification without using the decision threshold is based on the nearest-neighbor algorithm. The distances between the sequence of an individual who needs to be re-identified and all the sequences captured in another location are classified in increasing order. The closest sequence is chosen as the result of re-identification. If this sequence corresponds to the same person in the comparison, we obtain a true re-identification. By using this method of evaluation, we obtain similar results to those previously obtained.

Fig. 11 shows an example of the top five matching sequences for several query passages. The query passages are shown in the left column, while the remaining columns present the closest sequences ordered from left to right. The red box highlights the candidate sequence corresponding to the same person of the query. In this figure, the two cases of the first and second rows correspond to a true re-identification, while the third row falls in a false re-identification (the correct match is not the nearest sequence).

7. 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. Our proposed approach is based on the spectral classification of the color-based signatures extracted from the detected person in each sequence. A new descriptor called "color-position" histogram combined with several invariant methods is proposed to characterize the silhouettes in static images and obtain robust signatures which are invariant to lighting conditions. In order to further improve the

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

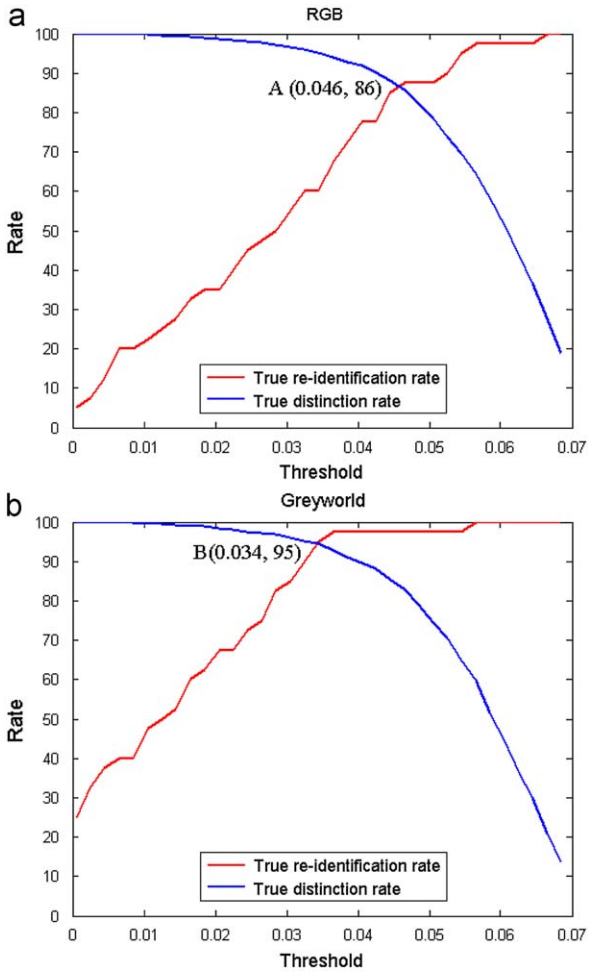

Fig. 9. Variation of TRR and TDR according to the settings of threshold. Two ETR points (A, B) corresponding to two color spaces (RGB and Greyworld, respectively) are presented. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
TRRs at the optimal points corresponding to RGB space and five color normalization procedures obtained by the proposed approach.

	TRRs at the optimal points (%)
RGB	86
Chromaticity space	65
Greyworld normalization	95
Comprehensive normalization	80
RGB-rank	88
Affine normalization	91

appearance-based model of an individual, many images of a video sequence should be exploited. Hence, an algorithm which is based on spectral analysis coupled with a specific SVM-based classification is applied to compare two sequences and make the final decision of re-identification.

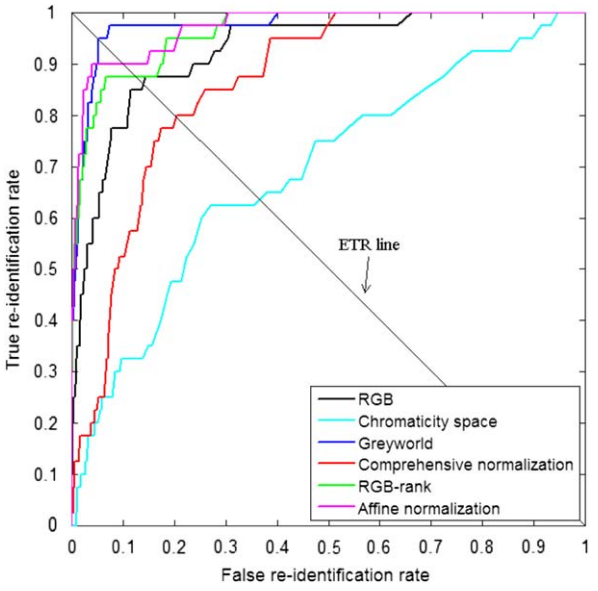

Fig. 10. ROC curves for comparing five invariant methods.

The global system was tested on a real and difficult dataset 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: 95% for the true re-identification rate and the distinction rate as well. These results are the fruit of the clever combination of the spectral analysis, the SVM method and the illuminant invariance of the color position silhouettes.

In order to further improve the performance of our system, the appearance-based signatures need to add more temporal and spatial information in order to be further discriminating among different people and to be unifying in order to make coherent classes with all the features belonging to the same person. An additional classification should be carried out in the case where the appearance of a moving person changes significantly due to occlusion, partial detection, etc. The other features, such as camera transition time, moving direction of the individual, biometrics features (face, gait), etc. should also be considered in order to improve the performance of the re-identification system, especially in the more challenging scenarios (multiple passages in front of cameras, many people wearing same color clothes, etc.).

More extensive evaluation also needs to be carried out. A good occasion will be to test it on people tracking in transport environment in the framework of the European BOSS project. 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.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

Fig. 11. Example of the top five matching sequences for several query passages.

Appendix A. Support vector machines

The SVMs were developed by Vapnik et al. [28]. They are based on the structural risk minimization principle from statistical learning theory. SVMs express predictions in terms of a linear combination of kernel functions centered on a subset of the training data, known as support vectors.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

Given the training data $(\mathbf{x}_i, y_i)$, $i = \{1, \ldots, m\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$, a SVM maps the input vector x into a high-dimensional feature space H through some mapping functions $\phi : \mathbb{R}^n \to H$, and builds an optimal separating hyperplane in this space. The mapping $\phi(\cdot)$ is performed by a kernel function $K(\cdot, \cdot)$ that defines an inner product in H. The kernel function maps the input space into a high dimensional Euclidean space and this kernel trick enables non-linear classification. A typical kernel is Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$. The separating hyperplane given by a SVM is $\mathbf{w} \cdot \phi(\mathbf{w}) + b = 0$. The optimal hyperplane is characterized by the maximal distance to the closest training data. The margin is inversely proportional to the norm of $\mathbf{w}$. Thus, computing this hyperplane is equivalent to minimize the following optimization problem [29]:

$$\mathcal{V}(\mathbf{w}, b, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{m} \xi_i\right) \tag{A.1}$$

where the constraint $\forall_{i=1}^{m} : y_i[\mathbf{w} \cdot \phi(\mathbf{x}_i) + b] \geq 1 - \xi_i, \xi_i \geq 0$ requires that all training examples are correctly classified up to some slack ξ and C is a parameter allowing trading-off between training errors and model complexity. This optimization is a convex quadratic programming problem. Its whole dual [28] is to maximize the following optimization problem:

$$\mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{w}_i, \mathbf{w}_j) \tag{A.2}$$

subject to $\forall_{i=1}^{m} : 0 \leq \alpha_i \leq C$, $\sum_{i=1}^{m} y_i \alpha_i = 0$.

The optimal solution α^* specifies the coefficients for the optimal hyperplane $\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \phi(\mathbf{x}_i)$ and defines the subset SV of all support vectors. An example $\mathbf{x}_i$ of the training set is a SV if $\alpha_i^* \geq 0$ in the optimal solution. The support vectors subset gives the binary decision function h:

$$h(\mathbf{x}) = \text{sign}(f(\mathbf{x})), \quad f(x) = \sum_{i \in SV} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b^* \tag{A.3}$$

where the threshold b^* is computed via the unbounded support vectors [28] (i.e. $0 < \alpha_i^* < C$).

An efficient algorithm SMO [30] and many refinements [31] were proposed to solve dual problem. SVM are powerful classifiers having high generalization abilities, but the decision function build by SVM has a complexity that increases with training set size. Moreover, high dimensional spaces are sensitive to the curse of dimensionality and scalar products quickly become hard to compute. One way to cope with these problems is to reduce the size of the input space by dimensionality reduction. Moreover, dimensionality reduction can ease the learning process and improve generalization abilities.

References

[1] ⟨http://www.celtic-boss.org/⟩.
[2] H. Ueda, T. Miyatake, S. Yoshizawa, An interavtive natural motion picture dedicated multimedia authoring system, in: CHI '91 Conference Proceedings, ACM Press, New York, 1991, pp. 343–350.
[3] A. Ferman, A. Tekalp, Multiscale content extraction and representation for video indexing, in: Multimedia Storage and Archiving Systems II, Proceedings of the SPIE, vol. 3229, 1997, pp. 23–31.
[4] X. Sun, M. Kankanhalli, Y. Zhu, J. Wu, Content-based representative frame extraction for digital video, in: IEEE Conference of Multimedia Computing and Systems, 1998, pp. 190–193.
[5] A. Girgensohn, J. Boreczky, Time-constrained keyframe selection technique, Multimedia Tools and Applications 11 (3) (2000) 347–358.
[6] Y. Yu, D. Harwood, K. Yoon, L. Davis, Human appearance modeling for matching across video sequences, Machine Vision and Applications 18 (3) (2007) 139–149.
[7] A. Ferman, S. Krishnamachari, A. Tekalp, M. Abdel-Mottaleb, R. Mehrotra, Group-of-frames/pictures color histogram descriptors for multimedia applications, in: Proceedings of the IEEE International Conference on Image Processing, 2000, pp. 65–68.
[8] A. Ferman, A. Tekalp, R. Mehrotra, Robust color histogram descriptors for video segment retrieval and identification, IEEE Transactions on Image Processing 11 (5) (2002) 497–508.
[9] T. Leclercq, L. Khoudour, L. Macaire, J.-G. Postaire, Compact color video signature by principal component analysis, in: Proceedings of the International Conference on Computer Vision and Graphics, 2004, pp. 814–819.
[10] N. Gheissari, T. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA, 2006, pp. 1528–1535.
[11] F. Porikli, O. Tuzel, Human body tracking by adaptive background models and mean-shift analysis, in: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, March 2003.
[12] D. Hall, J. Nascimento, P. Ribeiro, E. Andrade, P. Moreno, S. Pesnel, T. List, R. Emonet, R. Fisher, J. Victor, J. Crowley, Comparison of target detection algorithms using adaptive background models, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 113–120.
[13] O. Javed, Z. Rasheed, K. Shafique, M. Shah, Tracking across multiple cameras with disjoint views, in: Ninth IEEE International Conference on Computer Vision, 2003.
[14] K. Yoon, D. Harwood, L. Davis, Appearance-based person recognition using color/path-length profile, Journal of Visual Communication and Image Representation 17 (3) (2006) 605–622.
[15] S. Birchfield, S. Rangarajan, Spatiograms versus histograms for region-based tracking, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, 2005, pp. 1158–1163.
[16] G. Buchsbaum, A spatial processor model for object color perception, Journal of the Franklin Institute 310 (1) (1980) 1–26.
[17] G.D. Finlayson, B. Schiele, J. Crowley, Comprehensive color image normalization, in: Proceedings of the Fifth European Conference on Computer Vision, 1998, pp. 475–490.
[18] G. Finlayson, S. Hordley, G. Schaefer, G. Yun Tian, Illuminant and device invariant colour using histogram equalisation, Pattern Recognition 38 (2) (2005) 179–190.
[19] Y. Bengio, O. Delalleau, N.L. Roux, J.-F. Paiement, P. Vincent, M. Ouimet, Spectral dimensionality reduction, in: I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), Feature Extraction, Foundations and Applications, Studies in Fuzziness and Soft Computing, vol. 207, Springer, Berlin, 2006, pp. 519–550.
[20] L. Saul, K. Weinberger, J. Ham, F. Sha, D. Lee, Spectral methods for dimensionality reduction, in: O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT Press, Cambridge, MA, 2006, pp. 279–294.
[21] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[22] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[23] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.
[24] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators, in: Advances in Neural Information Processing Systems, 2005, pp. 955–962.
[25] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2001, pp. 849–856.
[26] U. von Luxburg, A tutorial on spectral clustering, Statistics and Computing 17 (4) (2007) 395–416.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005

[27] U. von Luxburg, O. Bousquet, M. Belkin, Consistency of spectral clustering, Technical Report 134, Max Planck Institute for Biological Cybernetics, 2004.
[28] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[29] J. Shawe-Taylor, N. Cristianini, Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, MA, 2000.
[30] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 185–208.
[31] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, Software Available at: 〈http://www.csie.ntu.edu.tw/∼cjlin/libsvm〉, 2001.

Please cite this article as: D.-N. Truong Cong, et al., People re-identification by spectral classification of silhouettes, Signal Process. (2009), doi:10.1016/j.sigpro.2009.09.005