People Detection and Re-Identification in Complex Environments

Dung-Nghi TRUONG CONG†, Member, Louahdi KOUDOUR†, Catherine ACHARD††, and Lounis DOUADI†, Nonmembers

SUMMARY This paper presents an automatic system for detecting and re-identifying people moving in different sites with non-overlapping views. We first propose an automatic process for silhouette extraction based on the combination of an adaptive background subtraction algorithm and a motion detection module. Such a combination takes advantage of both approaches and is able to tackle the problem of particular environments. The silhouette extraction results are then clustered based on their spatial belonging and colorimetric characteristics in order to preserve only the key regions that effectively represent the appearance of a person. The next important step consists in characterizing the extracted silhouettes by the appearance-based signatures. Our proposed descriptor, which includes both color and spatial feature of objects, leads to satisfying results compared to other descriptors in the literature. Since the passage of a person needs to be characterized by multiple frames, a large quantity of data has to be processed. Thus, a graph-based algorithm is used to realize the comparison of passages of people in front of cameras and to make the final decision of re-identification. The global system is tested on two real and difficult data sets recorded in very different environments. The experimental results show that our proposed system leads to very satisfactory results.

1. Introduction

In recent years, there has been a considerable interest in multi-camera systems for intelligent surveillance. More and more cameras with non-overlapping fields of view are introduced to provide coverage of wide areas in public places in order to meet requirements of security and monitoring tasks, such as detection of aggressions, loitering, or intrusion in forbidden areas. Existing surveillance systems rely on human observation of video streams for high-level classification and recognition. However, a large number of cameras makes this solution inefficient and, in many cases, unfeasible. Although the basic imaging technologies for simple surveillance are available today, their reliable deployment in a large network is still ongoing research.

In this article, we tackle the problem of people detection and re-identification in the extremely complex environments inside moving trains. The video sequences capturing moving people are analyzed in order to re-establish a match of the same person over different camera views located at different physical sites. In most cases, such a system relies on building 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. Kettner and Zabih [1] exploit the similarity of views of the person, as well as plausibility of transition time from one camera to another. Nakajima et al. [2] present a system which can recognize full-body people in indoor environments by using multi-class SVMs that were trained on color-based and shape-based features extracted from the silhouette. Javed et al. [3] use various features based on space-time (entry/exit locations, velocity, travel time) and appearance (color histogram). A probabilistic framework is developed to identify best matches. Bird et al. [4] detect loitering individuals by matching pedestrians intermittently spotted in the camera field of view over a long time. Snapshots of pedestrians are extracted and divided into thin horizontal slices. The feature vector is based on color in each slice and Linear Discriminant Analysis is used to reduce the dimension. Gheissari et al. [5] propose a temporal signature which is invariant to the position of the body and the dynamic appearance of clothing within a video shot. Wang et al. [6] represent objects using histograms of oriented gradients that incorporate detailed spatial distribution of the color of objects across different body parts. Yang et al. [7] propose 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.

In this article, we propose a system for automatically detecting and re-identifying people on-board trains whose synopsis is presented in Fig. 1. The proposed system consists of three main functions: silhouette extraction, appearance modeling for people characterization and people re-identification. Several originalities are presented in this paper: (i) in order to obtain pertinent results of moving object extraction, we propose to combine two well-known detection algorithms. As presented below, this combination allows us to take advantage of both methods by extracting the moving objects with highly precise silhouettes on the one hand and by managing important lighting changes on the other hand; (ii) the selection of well-extracted regions is also an important task that selects, among all the regions resulting from the detection, only those that really represent
silhouettes. In this work, we propose to use, in a particular way, a double criterion based on spatial and colorimetric coherence; (iii) in order to characterize each silhouette, a new descriptor, called SPAAtial and COlor DEscriptor (SPA-CODE), is introduced. It describes the color of the silhouette as well as the spatial arrangement of colors. As proved below on two different datasets, it leads to better results than other descriptors of the literature; (iv) the latest originality is introduced in the final step of people re-identification: in order to deal with the big quantity of data, we propose the use of the graph-based approach for dimensionality reduction. This allows us to obtain a robust distance between passages of people in front of cameras and to make the final decision of re-identification.

The outline of the paper is as follows: after this introduction, we present in Sect. 2 the proposed approach to extract moving regions. The selection of the extraction results, which keeps only the regions corresponding to silhouettes, is detailed in Sect. 3. Section 4 describes how the invariant signature SPACODE of a detected person is generated. In Sect. 5, we explain how we adapt the graph-based approach to our problematic of re-identification. Section 6 presents global performance of our system on two real datasets. Finally, in Sect. 7, conclusions and important short-term perspectives are given.

2. Moving Object Detection in Noisy Environments

Moving object detection, which provides a classification of the pixels into either foreground (moving objects) or background, is a critical task in many computer-vision applications. A common approach to detect the moving objects is background subtraction, where each new frame is compared to the estimated background model. Many approaches have been proposed in literature for building a robust background model. One of the simplest approaches consists in modeling each pixel intensity with a single Gaussian distribution [8]. However, such a model is unsuitable for noisy sequences and multi-modal scenes. More complex models can be based on a mixture of Gaussians [9], or a probability density function estimated by the kernel function [10]. The background can also be represented by a group of clusters which represent a compressed form of background model [11], [12]. Although the existing techniques have undeniable advantages, moving object detection in complex environments is still far from being completely solved.

As mentioned in the introduction, our research aims to set up an onboard surveillance system. Indeed, the complex environments inside a moving train make the moving objects detection task extremely difficult. Moreover, in our application, since people normally move towards the camera, an inflexible update of the background model leads to the inclusion of people in the background and the loss of significant parts of their silhouette. Therefore, in order to deal with such particular problems, we propose an approach based on an adaptive background subtraction algorithm coupled with a motion detection module. The synopsis of the proposed approach is shown in Fig. 2.

2.1 Moving Object Detection by Using Gaussian Mixture Models (GMM)

The GMM method for background subtraction [9] consists in estimating a density function for each pixel. The pixel distribution is modeled as a mixture of $N_G$ gaussians. The probability of occurrence of a color at the given pixel $I'(p)$ is estimated as:

$$p(I'(p) | I_p) = \frac{1}{N_G} \sum_{i=1}^{N_G} w_i \eta \left( I'(p) | \mu'_i, \Sigma'_i \right)$$

(1)

where $w_i$ is the mixing weight of the $i$th component at instant $t \left( \sum_{i=1}^{N_G} w_i = 1 \right)$; $\mu'_i$ and $\Sigma'_i$ are the estimates of the mean and the covariance matrix that describe the $i$th Gaussian component $\eta \left( I'(p) | \mu'_i, \Sigma'_i \right)$. Assuming that the three color components are independent and have the same variances, the covariance matrix is of the form $\Sigma'_i = \sigma'_I \mathbf{I}$.

The current pixel $I'(p)$ is associated with Gaussian component $k$ if $||I'(p) - \mu'_k|| < S_d \sigma'_I$, where $S_d$ is a multiplying coefficient of the standard deviation of a given Gaussian. The value of $S_d$ generally lies between 2.5 and 4, depending on the variation of lighting condition of the scene. The parameters of the matched component $k$ are then updated as follows:
\[
\mu_k' \leftarrow \left(1 - \frac{\alpha(p)}{w_k'}\right)\mu_k' - \frac{\alpha(p)}{w_k'}f(p),
\]
\[
\sigma_k'^2 \leftarrow \left(1 - \frac{\alpha(p)}{w_k'}\right)^2\sigma_k'^2 + \frac{\alpha(p)}{w_k'}\left(f(p) - \mu_k'ight)^2.
\]
\[
w_k' \leftarrow (1 - \alpha(p))w_k' + \alpha(p)
\]
where \(\alpha(p)\) is the update coefficient of pixel \(p\). Note that the update coefficient varies from pixel to pixel. An update matrix \(A\) that defines the update coefficient of each pixel will be estimated at the final stage of the moving object detection process (Section 2.3).

For the other components that do not satisfy the above condition, their weights are adjusted with:
\[
w_k' \leftarrow (1 - \alpha(p))w_k' - \alpha(p)
\]

If no matched component can be found, the component with the least weight is replaced by a new component with mean \(f(p)\), a large initial variance \(\sigma_0\) and a small weight \(w_0\).

In order to determine whether \(f(p)\) is a foreground pixel, all components are first ranked according to their values of \(\frac{w_k'}{\sigma_k'^2}\). High-rank components, which have low variances and high probabilities, are typical characteristics of background. The first \(C\) components describing the background are then selected by the following criterion:
\[
C = \text{argmin}_k \left\{ \sum_{i=1}^{C} w_i' > S_B \right\}
\]
where \(S_B\) is the rank threshold which measure the minimum portion of the components that should be accounted for the background. The more complex the background motion, the more the number of gaussians needed and the higher the value of \(S_B\).

Pixel \(f(p)\) is declared as a background pixel if \(f(p)\) is associated with one of the background components. Otherwise, it is detected as a foreground pixel.

Figure 3(b) illustrates an example of the moving object detection result by modeling the background with a mixture of Gaussians. A shadow removal step is also carried out by applying the algorithm proposed by Porikli and Tuzel [13] in order to improve the accuracy of the moving object extraction process. Connected component labeling [14] is then performed in order to remove noise and identify \(n\) disconnected regions \(B_1', B_2', \ldots, B_n'\).

For this difficult case, the results obtained by background subtraction using GMM and shadow removal are still noisy due to sudden illumination changes. Therefore, a fast-adapting algorithm could be combined with the GMM method in order to produce better localizations of moving objects and to eliminate all the regions that do not correspond to the foreground.

2.2 Moving Regions Detection

The objective of this step is to maintain the regions belonging to real moving objects and to eliminate noise and false detections. In addition, an update matrix that defines an adaptive mode for updating the background model is produced in order to provide additional robustness to the silhouette extraction process.

The motion detection module exploits the difference between three consecutive images. This technique has the advantage of requiring very few resources and adapting very quickly to changes in the background. The binary motion detection mask is defined by:
\[
M'(p) = \begin{cases} 1 & \left| f(p) - f(p-1) - \mu_1 \right| > S_M \cdot \sigma_1 \\ \cup & \left| f(p-2) - f(p-1) - \mu_2 \right| > S_M \cdot \sigma_2 \end{cases}
\]
where \(\mu_1\) and \(\sigma_1\) are the mean and the standard deviation of \(f'(p)\) and \(S_M\) is a threshold of the difference between \(f'\) and \(f''\). The optimal threshold \(S_M\) for binarization can be defined easily after several tests.

Figure 3(c) illustrates the motion detection result in which the edges of moving objects are detected. This detection result is then used to search among the \(n\) regions \(B_1', B_2', \ldots, B_n'\) obtained by the background subtraction and shadow removal module, those that correspond to the real moving objects (the silhouettes in our application). To do this, the number of pixels in motion of each region \(B_k'\) is estimated by:
\[
N_{B_k'} = \left\{ p : p \in B_k', M'(p) = 1 \right\}
\]

All regions whose number of pixels in motion is inferior to a threshold are eliminated. Let \(R' = \{r_1', r_2', \ldots, r_N'\}\) be

![Fig. 3](image-url)
the set composed by the \( N \) retained regions that are considered as the regions in motion (i.e. the regions corresponding to the moving objects). Figure 3(d) illustrates the result of this step.

### 2.3 Adaptive Update of the Background Model

The update matrix that defines the update coefficient of the Gaussian mixture of each pixel is re-estimated at this step. Initially, the update matrix is defined as a constant matrix \( \mathbf{A} = [\alpha(p)] \), \( \alpha(p) = M \) \( \forall p \in I \), where \( M \) is a constant. Then, the coefficients of this matrix are re-estimated as follows:

\[
\alpha(p) = \begin{cases} 
m & \text{if } p \text{ is detected as a pixel in motion} \\
M & \text{otherwise} 
\end{cases}
\]

where \( m << M \).

Thus, the pixels that are considered corresponding to the moving object will be slightly updated. This avoids integrating the silhouette of the people in the background model when they move towards the camera.

Although the proposed algorithm is quite simple, it is able to tackle the problems of difficult environments by extracting the moving objects with highly precise silhouettes thanks to the background subtraction algorithm based on GMM coupled with an adaptive update of the background model, and by managing important illumination changes with the moving region detection module. In this particularly difficult application inside a moving train where lighting is not stable and distracting motions are very numerous, it is indispensable to obtain relevant results.

### 3. Classification of Silhouette Extraction Results

In the previous section, we presented our approach to silhouette extraction that combines an adaptive background subtraction algorithm and a motion detection module. However, the silhouette extraction process still produces noisy results. Therefore, it is necessary to select, among all the regions \( r'_i \) obtained by the extraction process, the key-regions that correspond to silhouettes and effectively represent the appearance of a person. This step is particularly important, since it is indispensable to characterize only the regions corresponding to the silhouettes to obtain pertinent re-identification. The proposed algorithm for selecting the key-regions is based on both spatial and colorimetric coherence.

#### 3.1 Grouping Regions According to Their Spatial Coherence

The first step is to combine temporal regions of the same object by using the spatial information: it is reasonable to assume that two regions from two consecutive frames whose centers of gravity are close probably belong to the same object. Let \( \mathbf{R} = \{R^1, R^2, \ldots, R^T\} \) be all regions detected in \( T \) frames, with \( R^t = \{r^t_1, r^t_2, \ldots, r^t_N\} \). The regions are grouped into classes according to the algorithm 1. where \( C(r^t_j) \) is the class of region \( r^t_j \) and \( \text{dist}(r^t_{j-1}, r^t_j) \) is the distance between the centers of gravity of regions \( r^t_{j-1} \) and \( r^t_j \). \( S_r \) is a threshold for classifying regions.

A result obtained with this spatial coherence criterion is presented in Fig. 4. For simplicity of representation, each class is illustrated by one frame.

#### Algorithm 1: Classification based on spatial coherence

Initialization: \( \forall j \in [1, N] \), \( C(r^0_j) \leftarrow j \).

while \( t \leq T \) do

for each region \( r^t_j \) do

if \( \exists i \) \( \text{dist}(r^t_{i-1}, r^t_j) < S_r \), then

\( C(r^t_j) = C(r^t_{i-1}) \)

else

\( C(r^t_j) = \text{new class} \)

\( t = t + 1 \)

end

end

end
Algorithm 2: Classification based on colorimetric characteristics

Initialization: $G_1 \leftarrow C_1$, $n = 1$

for each class $C_k$ do
  \[ j = \text{argmin}_j \{\text{dist}(C_k, G_j)\} \]
  if $\text{dist}(C_k, G_j) < S_C$ then
    $G_j \leftarrow C_k$
  else
    create a new group: $n = n + 1$; $G_n \leftarrow C_k$

\[ \text{dist}(C_k, G_i) = \min \{\text{dist}(C_k, C_{ib})\}, b = 1 \ldots B \] (8)

where $C_{ib}$ is a class belonging to group $G_i$. The distance between two classes $\text{dist}(C_p, C_q)$ is defined by:

\[ \text{dist}(C_p, C_q) = \frac{1}{P \times Q} \sum_{i=1}^{P} \sum_{j=1}^{Q} d(X_{pi}, X_{qj}) \] (9)

where $X_{pi}$ is the descriptor of region $i$ of class $C_p$, $P$ and $Q$ are the numbers of regions in classes $p$ and $q$ respectively.

We assume that the most stable sets (i.e. the sets that have a great number of selected regions) correspond to the silhouette of the moving person. We call it the set of silhouettes and use it to characterize the passage of people in front of a camera.

4. Invariant Appearance Model

In the previous section, we presented the specific algorithm for silhouette extraction. The next important step of our system is to describe the appearance of moving people detected in video sequence, where geometric deformations and illumination variations induce large changes in appearance. Thus, the appearance model should be robust to illumination, viewpoint changes, as well as object deformations.

In order to obtain an appearance model which is robust to illumination, a first step of color normalization procedure has to be carried out. Many methods have been proposed in the literature and we tested most of them. In this paper, we focus on the three invariances that led to better results:

- Greyworld normalization [16] is derived from the RGB space by dividing the pixel value by the average of the image (or in the area corresponding to the moving person in our case) for each channel:

\[ I_k^* = \frac{I_k - \text{mean}(I_k)}{\text{std}(I_k)} \] (10)

where $I_k$ is the color value of channel $k$.

- Normalization using histogram equalization [17] is based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging illuminations. The rank measure for level $l$ and channel $k$ is obtained with:

\[ M_k(i) = \sum_{u=0}^{i} H_k(u) \big/ \sum_{u=0}^{Nb} H_k(u) \] (11)

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

- Affine normalization is based on the illumination change model [18] in which the sensor responses under a pair of illuminants are related by a diagonal matrix and translation transform:

\[ \Pi^1 = \mathbf{D} \mathbf{I}^2 + \mathbf{T} \] (12)

where the superscripts 1 and 2 represent the pair of illuminants, $\mathbf{D}$ is a diagonal matrix and $\mathbf{T}$ is the translation vector.

Affine normalization is thus defined by:

\[ I^*_k = \frac{I_k - \text{mean}(I_k)}{\text{std}(I_k)} \] (13)

Since the appearance of people is dominated by their clothes, color-based models are suitable for their description. The most widely used feature is color histograms [3] that are resistant to deformable shapes and invariant to scale by normalization. The main drawback of color histograms is the lack of spatial information, which can be retained by using spatiograms [19]. The second-order spatiogram, which is a generalization of histogram, includes the spatial mean and covariance of pixels belonging to each histogram bin. Some other approaches for building appearance models have also been proposed in literature, such as color/path-length descriptor [20], histograms of oriented gradients in the log-RGB color space [6], statistical color-position feature [21].

In this article, we propose a new appearance model, called SPACODE (SPAtial and COlor DEscriptor), based on the color and spatial distribution of pixels inside the silhouette. Each pixel is represented by a feature vector $x = [c, l]$, where $l$ is the normalized length between the top point of head and the current pixel, and $c$ is a vector of $k$ color components derived from the color normalization space. Here, we choose the top point of head as the base point since it is easy to detect from the extracted silhouettes and relatively stable to movement.

The probability density of feature vector $x$ is estimated by using an adaptive mixture of $N_c$ Gaussians:

\[ f(x) = \sum_{i=1}^{N_c} \alpha_i \eta(x|\mu_i, \Sigma_i) \] (14)

where $\alpha_i$ is the mixing weight of the $i^{th}$ component, $\eta(x|\mu_i, \Sigma_i)$ is the Probability Density Function of normal distribution with mean vector $\mu_i$ and covariance matrix $\Sigma_i$:

\[ \eta(x|\mu_i, \Sigma_i) = \frac{\exp \left( -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right) }{\sqrt{(2\pi)^{k+1} |\Sigma_i|}} \] (15)

In many systems, Gaussian mixture model is effectively estimated using the Expectation Maximization (EM)
algorithm [22]. However, there are some existing problems in using EM. Firstly, the number of components must be predefined, while we wish to adapt the model in response to appearance changes. Secondly, a large number of iterations are required, which makes the algorithm computationally expensive. Another problem is that the convergence behavior of the EM algorithm is quite sensitive to the starting point and a convergence to a local minimum or a saddle point might occur.

Therefore, in order to avoid such drawbacks, we propose an algorithm based on the technique of splitting and merging components [23], [24] in order to build up the SPACODE descriptor. The main contribution of this approach is that it is able to generate an appropriate representation of the silhouette by adapting the number of Gaussians to the major regions of the silhouette. Furthermore, it is computationally manageable.

Given a set of $N$ feature vectors $x_n$ derived from the current silhouette. The model is initialized with a single component whose mean vector $\mu_0 = \frac{1}{N}\sum_{n=1}^{N} x_n$ and covariance matrix $\Sigma_0 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_0)^T (x_n - \mu_0)$. The mixture model is then adapted by iteratively applying the splitting and merging procedure.

- **Splitting components:**

  Given the current set of components, for each component $n$, let $\Sigma_n = V_n \Lambda_n V_n^T$ be the spectral decomposition of the covariance matrix $\Sigma_n$, where $\Lambda_n = \text{diag} (\lambda_n^1, \ldots, \lambda_n^{n+1})$ is a diagonal matrix containing the eigenvalues of $\Sigma_n$, $V_n$ is the matrix consisting of the eigenvectors of $\Sigma_n$, one eigenvector per column. Component $n$ is decomposed into two new component $n_1$ and $n_2$ if its maximum eigenvalue is above a threshold:

  $$\lambda_n^{\text{max}} > T_S$$

  Here, $T_S$ is a threshold defining the largest variance of components. In our algorithm, the threshold $T_S$ is chosen as $T_S = p_v \lambda_0^{\text{max}}$, where $\lambda_0^{\text{max}}$ is the maximum eigenvalue of the covariance matrix $\Sigma_0$ of the initialization component, $p_v$ is a predefined percentage value modifying the variance of component ($p_v$ is set to $p_v = 20\%$ in our experiments).

  Thus, component $n$ which satisfies the above condition is split into two new components $n_1$ and $n_2$. Let $v_n^{\text{max}}$ be the eigenvector corresponding to the maximum eigenvalue $\lambda_n^{\text{max}}$ of the covariance matrix $\Sigma_n$. The two subsets of pixels re-assigned to these two new components are defined by the separating plane which is perpendicular to $v_n^{\text{max}}$ (Figure 5):

  $$C_{n_1} = \{ x \in C_n | (x - \mu_n)^T v_n^{\text{max}} \geq 0 \}$$
  $$C_{n_2} = \{ x \in C_n | (x - \mu_n)^T v_n^{\text{max}} < 0 \}$$  

  where $C_n$, $C_{n_1}$ and $C_{n_2}$ are the sets of feature vectors belonging to components $n$, $n_1$ and $n_2$ respectively.

  The parameters of both components are re-estimated based on the statistics of their assigned pixels:

  - **Merging components:**

    The criteria for merging two components are based on the covariances of these two components and the distances between them. Let $M_1 (\mu_2) = \sqrt{(\mu_2 - \mu_1)^T \Sigma_1^{-1} (\mu_2 - \mu_1)}$ be the Mahalanobis distance of mean vector $\mu_2$ of the second component from the first component with mean vector $\mu_1$ and covariance matrix $\Sigma_1$. Let $M_2 (\mu_1) = \sqrt{(\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2)}$ be the Mahalanobis distance of $\mu_1$ from the second component. Two components are considered suitable for merging if these two Mahalanobis distances are lower than a given threshold:

    $$M_1 (\mu_2) < T_m \land M_2 (\mu_1) < T_m$$

  where $T_m$ is a predefined threshold for merging.

  The parameters of the new merged Gaussian are calculated as:

  $$\alpha_{\text{new}} = \alpha_1 + \alpha_2$$
  $$\mu_{\text{new}} = \alpha_1 \mu_1 + \alpha_2 \mu_2$$
  $$\Sigma_{\text{new}} = \frac{1}{n_{\text{new}}} \sum_{i=1}^{n_{\text{new}}} \alpha_i \left( \Sigma_i + (\mu_i - \mu_{\text{new}})(\mu_i - \mu_{\text{new}})^T \right)$$

  These two procedures are repeated until reaching convergence. The silhouette of a person is thus represented by the SPACODE containing $N_c$ components of the mixture. Figure 6 illustrates an example of SPACODE consisted of 5 components which are represented in the $l$-marginal. The color of each Gaussian in figure 6b presents the mean color.
of the corresponding component.

The distance (dissimilarity) between two silhouettes is computed by first re-calculating the probability density function \( f(x) \) corresponding to each silhouette (Eq. 14), then using the normalized L2 distance as the distance between two probability density functions:

\[
d(X_1, X_2) = \sum_x \left( f'_1(x) - f'_2(x) \right)^2
\]

where \( f'(x) = f(x) / \sqrt{\sum_x f(x)^2} \).

5. Graph-Based Approach for People Re-Identification

5.1 Graph Notion and Random Walks View

In the previous section, we presented the invariant appearance-based signature extracted from each silhouette. A set of signatures that characterizes each passage of an individual in front of a camera is the input data of the people re-identification process. In this section, 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 \( X = \{X_1, X_2, \ldots, X_m\} \) extracted from \( m \) silhouettes belonging to two passages in front of cameras, this set is associated to a complete neighborhood graph \( G = (V, E) \) where each data point \( X_i \) 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 \( S_{ij} = \exp \left( -\frac{d(X_i, X_j)^2}{\omega^2} \right) \). Here, \( d(X_i, X_j) \) is the distance between two signatures and the parameter \( \omega \) is chosen as \( \omega = \text{mean} \left[ d \left( X_i, X_j \right) \right], \forall i, j = 1, \ldots, m \) \((i \neq j)\). Matrix \( S = [S_{ij}]_{i,j=1,\ldots,m} \) is called the similarity matrix. Let \( D \) denote the diagonal matrix with elements \( D_{ii} = \sum_i S_{ij} \) 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 \( P = \left[ P_{ij} \right]_{i,j=1,\ldots,m} \) of the random walk is thus defined by:

\[
P = D^{-1}S
\]

The set of the eigenvalues of \( P \), \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \), is usually called the spectrum of \( P \) (or the spectrum of the associated graph \( G \)). Since \( P \) is a right stochastic matrix and \( P_{ij} > 0, \forall i,j=1,\ldots,m \), the first eigenvalue \( \lambda_1 = 1 \) with the corresponding eigenvector \( \gamma_1 = [1, 1, \ldots, 1]^T / \sqrt{m} \). 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 \( h : X \rightarrow \mathbf{u}_i = [\gamma_1(i), \ldots, \gamma_q(i)] \) where \( \gamma_k(i) \) is the \( k \)th coordinate of eigenvector \( \gamma_k \).

Coming back to our problem, the set of signatures belonging to two passages is characterized by transition matrix \( P \) of the random walks on graph \( G \). Note that the data set is composed of two known clusters, one for each passage. Our problem consists in evaluating how separate these two clusters are; or, in other words, in measuring the similarity between two passages. 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 \( P \).

5.2 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 \( X = \{X_1, X_2, \ldots, X_m\} \) are ordered according to the cluster they belong to (i.e. the first points belong to passage \( a \) and the others to passage \( b \)). Since two clusters are infinitely apart, the similarity matrix is a block diagonal matrix \( \hat{S} = \begin{bmatrix} S^{(a)} & 0 \\ 0 & S^{(b)} \end{bmatrix} \) where \( S^{(a)} \) and \( S^{(b)} \) are the matrices of intra-cluster similarities of clusters \( a \) and \( b \) respectively.

Transition matrix \( \hat{P} \) 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 \( \hat{P} \) are 1 and the matrix containing the first two corresponding eigenvectors as columns is \( \hat{\Upsilon} = \begin{bmatrix} \gamma_1^{(a)} & 0 \\ 0 & \gamma_1^{(b)} \end{bmatrix} \).

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 \( (\mathbf{u}_a = [c(a)\ 0] \) for cluster \( a \) and \( \mathbf{u}_b = [0\ c(b)] \) for cluster \( b \). Therefore, the classification of points \( \mathbf{u}_a \) 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 [25] 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 passages, the larger the perturbations. The points \( X_i \) belonging to these two similar passages are not well separated. This leads to the fact that points \( \mathbf{u}_a \) 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 passages are quite different, we have a nearly ideal transition matrix \( \mathbf{P} \) whose off-diagonal blocks are approximately 0. Points \( \mathbf{u}_i \) 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 7 illustrates the variation of eigengap according to the similarity of two passages 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 passages. The first dataset consists of two passages of the same person captured in different locations (first row). The second and third data sets are composed respectively of two passages belonging to 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 are very distinct, and points \( \mathbf{u}_i \) are almost fixed for each cluster. The eigengap is large enough to ensure the stability of the eigenvectors.

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

6. Experimental Results

In this section, we perform the evaluation of the proposed system by using two real datasets which are described in the subsection 6.1. The experiments are carried out for double purposes. We first evaluate the silhouette extraction algorithm on the sequences in presence of fast illumination variations and non-static background. The second stage is for evaluating the total system performance of people re-identification.

6.1 Datasets for the Evaluation

As mentioned above, our research aims to set up an on-board surveillance system that is able to re-identify a person captured by different cameras. Before collecting the on-board dataset, another dataset acquired in INRETS premises was first collected for the evaluation of our algorithms. It contains passages of 40 people captured by two cameras. We have chosen two different locations (indoors in a hall near windows and outdoors with natural light) to set up those two cameras. The second dataset is collected in the framework of the BOSS European project (on BOard wireless Secured video Surveillance) [26]. This dataset contains passages of 35 people captured by two cameras installed on board a moving train. This dataset is really difficult for both silhouette extraction and people re-identification tasks, since these two cameras are set up with different angles and what is captured by the video system is influenced by many factors, such as fast illumination variations and non-static background due to the movement of the train, reflections, vibrations...

Figure 8 illustrates these two datasets. In this figure, we notice that the color appearance is very different according to the location of the person.

Figure 9 illustrates some images of the second dataset corresponding to one passage of a person in the corridor of the train. We note that the lighting conditions are very different even for a single passage.
6.2 Evaluation of Silhouette Extraction

In this section, we demonstrate the effectiveness of the proposed method for silhouette extraction in comparison with the standard GMM algorithm. Figure 10 illustrates some results obtained from the difficult sequences in presence of fast illumination variations and non-static background due to the movement of the train. The first column presents the original images, the second column shows the results obtained by the standard GMM method, and the third row gives the results obtained by the proposed algorithm. Note that the same parameters (except the update coefficient since it is adapted in the proposed algorithm) are used for the GMM in both tested methods.

One can notice that the standard GMM method fails to deal with the irregular variations of the background, which results in many false positive pixels. Furthermore, since people normally move towards the camera, an inflexible update of the background model in the standard GMM method leads to the loss of significant parts of their silhouette.

In order to perform a quantitative analysis of the proposed approach, we have manually segmented 800 frames from the sequences of the first dataset and 1400 frames from the second dataset. The performances of the proposed algorithm are evaluated by using recall and precision measurements, where

\[
\text{recall} = \frac{\text{number of true foreground pixels detected}}{\text{number of true foreground pixels}}
\]

\[
\text{precision} = \frac{\text{number of true foreground pixels detected}}{\text{number of foreground pixels detected}}
\]

Table 1 presents the comparative results of silhouette extraction obtained from both datasets. In order to make a fair comparison, shadow removal and morphological operations are also used in the tests of standard GMM method. Clearly, the results demonstrate that the proposed approach obtains better performance in both terms of recall and precision.

Figure 11 shows the per-frame detection accuracy in terms of recall and precision obtained from the second dataset. One can notice that the extraction accuracy of our approach is consistently higher than the standard GMM method.

6.3 Evaluation of People Re-Identification

The experiments for evaluating the total system performance of people re-identification were conducted in two stages. Firstly, we evaluate the proposed approach for peo-
ple re-identification by using ground truth silhouettes. The second series of experiments analyze the influence of automatic silhouette extraction on the results of re-identification.

In order to validate the proposed signature SPACODE, we compare it with the three signatures existing in literature:

- Color histograms with 8 bins per color channel. Histogram intersection that measures the similarity between two histograms is used to construct the weighed graph.
- Spatiograms with 8 bins for each color channel. The similarity measure proposed in [19] is used to construct the graph.
- Color/path-length descriptor with 8 bins per color channel and 8 bins for the path-length feature. The similarity between two signatures is computed by using Euclidean distance.

For each query passage in front of one camera, the similarities between the query passage and each of the candidate passages of the other camera are calculated. The most similar passage is chosen as the result of re-identification. If the chosen passage corresponds effectively to the same person as the query passage, we obtain a true re-identification ($T_i = 1$). Otherwise, we obtain a false re-identification ($T_i = 0$).

The true re-identification rate can be calculated by using the following definitions:

$$TRR = \frac{\sum_{i=1}^{N} (T_i = 1)}{N}$$

where $N$ is the number of people in the database.

Tables 2 and 3 report the true re-identification rate versus the descriptors obtained for two datasets. The best rates of 97.5% for the first dataset and 97.1% for the second are obtained by combining the proposed signature SPACODE and the invariant normalizations (Greyworld and affine normalization). We note that the performance of color histograms is the worst. The others lead to better results thanks to the additional spatial information.

The second test is to analyze the influence of automatic silhouette extraction on the results of re-identification. The comparative results of re-identification obtained by the ground truth silhouettes and the automatic detection are shown in Table 4. For the first dataset, automatic extraction does not degrade the performance of re-identification. For the second base, the re-identification rates are slightly degraded in comparison with those obtained by manual extraction due to the false detections in difficult conditions of this dataset. The best rate of 91.4% obtained by a completely automated process is encouraging considering the difficulties of the real data captured on board a train.

Concerning the processing time, on a 2 GHz Intel Core 2 Duo processor PC, the proposed approach for silhouette extraction and characterization can process about 3fps for a frame size of 360 by 288. The processing time for re-identification including one query passage and 40 candidate passages is about 6.7 seconds. The total processing time of the proposed system is fairly high for a real-time implementation. However, several improvements (such as opti-
mization of the software, more powerful computer utilization and specifically parallel architectures, ...) that will be carried out for the final system installed in the train can ensure a real-time functioning of the people detection/re-identification system.

7. Conclusion

In this paper, we have presented a surveillance system including multiple non-overlapping cameras and a specific algorithm to detect/re-identify people who appear and reappear at different times and across different cameras. The first originality introduced in this article is an automatic process for silhouette extraction based on the combination of an adaptive background subtraction algorithm and a motion detection module. Such a combination takes advantages of both approaches and is able to tackle the problem of particularly difficult environments, such as inside a moving train. As it is indispensable to characterize only the regions corresponding to the silhouettes, an original method is carried out to obtain key-regions. Regions extracted during the detection step are clustered based on their spatial belonging and colorimetric characteristics. The following step consists in characterizing the extracted silhouettes by the appearance-based signatures. The proposed descriptor called SPACODE, which includes both information on color and their spatial arrangement, leads to improve results compared to other descriptors of the literature. Since the passage of a person needs to be characterized by multiple frames, a large quantity of data has to be processed. Hence, rather than using simple reduction dimension methods, such as Principal Component Analysis, we introduced an algorithm, already employed in other applications, which is based on the random walk on the graph. It is used to measure the similarity between two passages with an important robustness and to make the final decision of re-identification.

The global system was tested on two real datasets. The first was captured in laboratory while the second comes from a real surveillance system including two sensors installed on board a moving train. The experimental results show that the proposed system leads to satisfactory results: 97.5% for the true re-identification rate for the first dataset and 91.4% for the second.

Since the re-identification problem has gained in interest in the last few years and since there is no common database for evaluation, a comparative study of similar works in literature for people re-identification is hard to carry out. Thus, in this article, we only compare the performance of our system with the others in literature based on the global re-identification rate. For instance, Gheissari et al. [5] have used a dataset with 44 people and have obtained a 60% true re-identification rate. Wang et al. [6] achieve a matching rate of 82% on a dataset containing 99 individuals. In the system proposed by Yu et al. [7], the accuracy is 95% for a 30 people dataset captured indoors. Nakajima et al. [2] have obtained a 100% accuracy but on a very reduced dataset (4 people). One can notice that our approach leads to better results if we take into account the large and diverse dataset used.

Furthermore, in order to make available to researchers a common database for comparison, the dataset captured by multiple cameras installed on board a moving train in the framework of the BOSS project can be found on the website [26]. We hope that it will be a reference database for evaluating similar works concerning people re-identification problem.

In order to further improve the performance of our system, the appearance-based features need the addition of more temporal information, such as camera transition time from one camera to another, spatial information, moving direction of people. These latters can help us deal with the more challenging scenarios (multiple passages in front of cameras, many people wearing clothes of the same color, occlusion, partial detection, etc).

References

[15] D.N. Truong Cong, L. Khoudour, C. Achard, and P. Phothisane,


Lounis Douadi  was born in 1978 in Algiers, Algeria. He received his M.S and Ph.D. degrees in robotics from the University of Montpellier, France in 2001 and 2006, respectively. He joined the French National Institute for Transportation and Safety Research (INRETS) in 2007. His research interest is in signal and image processing applied to perception systems for transportation. Previously, he has been working in the same field with applications in mobile robot localization.

Catherine Achard  was born in 1970. She received the Ph.D. degree from the University Blaise Pascal of Clermont-Ferrand in 1996, and the Leading research degree in 2007. She has been an associate professor at the University Pierre and Marie Curie since 1997. Her current research topics focus on video analysis, specifically on the study of human behavior.

Dung-Nghi Truong Cong  received the B.S. degree in telecommunications engineering from the Ho Chi Minh City University of Technology, Vietnam in 2004 and the M.S. degree in signal, image, speech and telecommunications from the Grenoble Institute of Technology, France in 2007. She is currently pursuing the Ph.D. degree at the University Pierre and Marie Curie (Univ. Paris 6), as a Fellow of the French National Institute for Transport and Safety Research (INRETS), France. Her research interests include image and video processing, computer vision, video surveillance, target detection and pattern recognition.