Intelligent distributed surveillance system for people re-identification in transport environment

Dung Nghi Truongcong, Louahdi Khoudour, Catherine Achard, Jean-Luc Bruyelle

Intelligent distributed surveillance system for people re-identification in transport environment

This paper presents a solution to track people across a network of cameras with disjoint fields of vision. Firstly, an appearance-based signature is extracted from each frame of the sequence characterizing the passage of a person. This feature, called "colour-position" signature, merges spatial and colour information for improved robustness. Moreover, an illuminant invariant procedure has been introduced to manage lighting and camera response changes. Secondly, the distance between two sequences is estimated thanks to dimensionality reduction techniques. Two methods are implemented and compared. The first one is the classical, linear, Principal Components Analysis, while the second one is a recent non-linear method called Laplacian Eigenmaps approach. Results from two databases acquired in difficult conditions, show that the Laplacian Eigenmaps method always yields the best results, which confirms the interest of non-linear approaches in such applications. Moreover, these results show the relevance of this approach combining appearance-based signature and dimensionality reduction technique within the scope of people tracking across distinct fields of vision.

Keywords: Surveillance systems; Smart camera; Person re-identification; Colour invariant; Principle Component Analysis; Laplacian Eigenmaps.

1 Introduction

Public transport operators are facing an increasing demand in efficiency and security, from both the general public and governments. An important part of the efforts deployed to meet these demands is the ever-increasing use of video surveillance cameras, in order to detect incidents and inform the staff without delay. A major drawback of this approach, however, is the very large number of cameras required to effectively monitor even a comparatively small network.

In recent years, image-processing solutions have been found to automatically detect incidents and make measurements on the video images recorded by the cameras, relieving the staff in the control room from much of the difficulty of finding out where interesting events are happening. However, the need remains to bring the data to a centralised computer, which leaves the need to install huge lengths of video cabling, increasing the cost, the complexity, and decreasing the adaptability of the video system.

In the framework of the BOSS European project (http://www.celtic-boss.org), INRETS (French National Institute for Transport and Safety Research) has devised and tested in real-life situations an architecture to address these problems. The general idea is to avoid sending many full-resolution, real-time images at the same time to the video processor, by placing the processing power close to the cameras themselves, and sending only the meaningful images to the control room. Until recently, computers and video grabbers were much too expensive to even dream of having multiple computers spread all over the network. But steadily decreasing costs already allow to realize such a network, although still at a cost.

In the project, we have developed a multi-camera vision system specified to meet key requirements of security and monitoring tasks on board trains. The diversity of these tasks implies significant processing power and highly versatile configurations.

These requirements are best met by a modular architecture based on localised image processing (at or near the camera), and distributed processing whereby cameras are connected to a local network, sending event information and video only upon detection of an event of interest.

In this paper, we propose a short description of the architecture developed, including two video sensors installed and a detailed description of a specific algorithm of video comparison between the two cameras. The objective of our work is to establish correspondence between observations of people who appear and reappear at different times and places. Our proposed approach relies on an appearance-based model combined with an algorithm that reduces the effective working space and realizes the comparison of the video sequences.

Therefore, the outline of the paper is as follows: Section 2 describes briefly the smart surveillance architecture. Section 3 describes how the invariant signature of a detected person is generated by the image processors of the associated cameras. In Section 4, after a few theoretical review of techniques for dimensionality reduction, we explain how we adapt the latter to our problem. Section 5 presents global performance results of our system on two real datasets. Finally, in Section 6, conclusions and important short-term perspectives are given.

2 Architecture of the network

Each camera is connected, via its own video cable, to a local image processor which is actually a miniature PC incorporating an image grabber and an Ethernet link. These processors are readily available on the market. Each processor, according to the model and the frame rate requirements, can process images coming from one to four cameras in sequence. These processors are small enough to be mounted in a false ceiling, or in an electric rail.

The image processors normally operate stand-alone, and only process images from the associated cameras, without sending or receiving any data. It is only in specific conditions that they communicate with the supervising computer.

All the processors are connected via Ethernet to the supervising computers which act as the HMI (Human Machine Interface) between the processors and the control room operators. The central computer also manages and backs up the data received from the processors, and resolves possible conflicts between data sent simultaneously by several processors.

Use of the network

As mentioned above, all the image processing is done locally, requiring minimal use of the network. But of course the network is still required in two cases:

- When an alarm is raised or when a special function is turned on (information from the local processors to the supervising computer or between two processors in case of information exchange).
- To set the processing parameters (parameters from the supervising computer to the local processors).

When an alarm is raised, the local processor tells the supervising computer that an incident is in progress. It can also, on request, send the image that triggered the alarm, in

order to give more information to the operators in the control room. It is also possible, on request from the supervising computer, to perform a special function or to download images from the camera for routine monitoring purpose when no alarm is raised.

Within the framework described in this paper, we develop a system able to re-identify a person who has appeared in the field of one camera and then reappears in front of another camera. **Figure 1** shows the structure of such a system. In normal mode, as soon as a person is detected by either camera, her signature is calculated, and added to the camera's database. Once the re-identification function is turned on (e.g. function "re-identification CAM2-CAM1"), all the signatures in the database of camera 1 are sent to camera 2. Then, camera 2 begins to perform the re-identification task: the new signature of each passage in front of camera 2 is first extracted and then compared with all the signatures received from camera 1. Conversely, the inverse operation is done in the "re-identification CAM1-CAM2" function. A specific algorithm, described in detail in this paper, is used for such video comparisons.

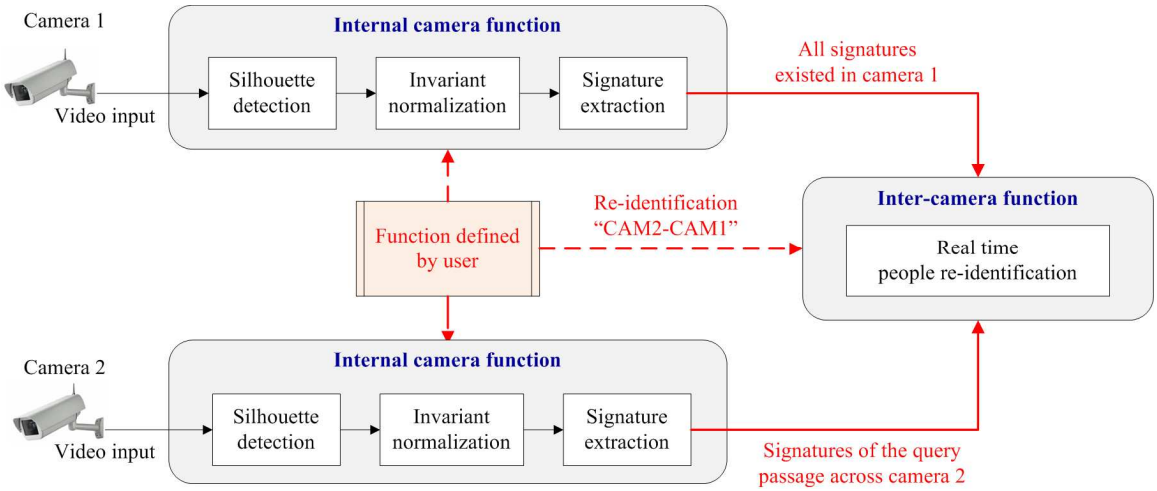


Figure 1: Structure of a system including two video sensors and its special function.

3 Appearance-based feature extraction

3.1 Silhouette extraction

A common approach for extracting moving objects (silhouettes in our case) from a video sequence is to use a background subtraction algorithm. The idea is to subtract the current frame from a reference background model. Since the natural scenes are not static in most situations, many models have been proposed for handling complex, non-static backgrounds, like the models presented by (Stauffer et al., 1999) or (Harville, 2002).

In our system, we use the codebook background subtraction algorithm (Kim et al., 2004) that adopts a quantization/clustering technique (Kohonen, 1998) to construct a background model. The algorithm encodes the time series of the observed colour values for each pixel location over time into codebooks. Samples at each pixel are clustered into a set of codewords. The similarity measure for this clustering is based on a colour distortion measure and a brightness range. In order to detect foreground objects in a new frame, each pixel colour and its brightness are compared with the codewords describing the background at this location. A pixel is classified as background if its colour distortion with respect to the codeword is lower than a given threshold, and its brightness lies within the brightness range of that codeword. Otherwise, it is classified

as foreground.

Figure 2 presents an example of silhouette extraction by using the codebook background subtraction algorithm. The background subtraction results have several local errors. Hence, in order to obtain more accurate silhouettes, morphological operators (erosion and dilation) are first applied to the binary images. Then, an algorithm called “Connected Component Labelling” (CCL) (Ballard and Brown, 1982) is used to group pixels into continuous regions based on pixel connectivity. The largest isolated region obtained by CCL algorithm is considered as the desired silhouette, while the other regions are removed from the background.

Figure 2: Silhouette extraction (left to right, top to bottom: original frame, background subtraction by codebook algorithm, result obtained by morphological operators, final silhouette after applying CCL algorithm).

3.2 Colour-based signature extraction

An important step of our system consists in extracting from each frame a robust signature characterizing the passage of a person. Since the appearance of people is dominated by their clothes, colour features are suitable for their description.

The most widely-used feature for describing the colour of objects is colour histograms (Javed et al. 2003, Nakajima et al., 2003) that are robust to deformable shapes and scale-invariant by normalization. The main drawback of colour histograms is the lack of spatial information. This leads to the fact that they cannot discriminate between appearances with the same colour distribution, but different colour structures. Several approaches have been proposed to include spatial information in the histogram

format. Nakajima et al. (2003) proposed a new histogram format by combining the colour histogram with the shape histogram calculated by counting pixels along rows and columns of the extracted images. Yoon et al. (2006) and Yu et al. (2007) presented the colour/path-length feature which includes some spatial information: each pixel inside the silhouette is represented by a feature vector $(\mathbf{x}, l)$, where $\mathbf{x}$ is the colour value and l is the length between an anchor point (the top of the head) and the pixel. The distribution of $p(\mathbf{x}, l)$ is then estimated with a 4D histogram. We can lastly cite spatiograms (Birchfield et al., 2005), 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.

We propose a new descriptor for static images called the "colour-position" signature (**Figure 3**). The idea of this signature is that the interested region (the silhouette in our case) can be horizontally decomposed in areas with homogeneous colour. Thus, for estimating this new descriptor, the mean colour of each row of the silhouette is calculated. The "colour-position" signature $\mathbf{S}$ is now composed of $N = p \times q$ values $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_p\}$, where p is the total number of equal parts of the silhouette and $\mathbf{s}_i$ is a vector of q colour components. The advantages of such a signature are its consideration of the spatial information, its simple estimation and its low memory consumption.

Figure 3: Colour-position signature (left to right: silhouette extraction, colour-position signature extraction, signature vector).

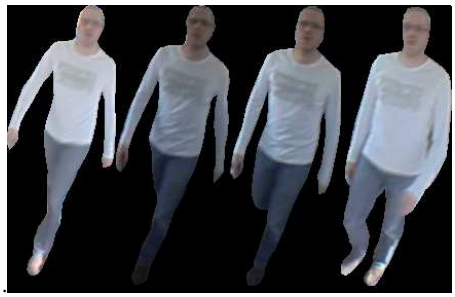
3.3 Invariant normalization

Since our surveillance system is set up inside an operating train, the quality of the video sequences acquired by cameras depends heavily on several factors, such as fast illumination variations, light reflections, vibrations, etc. One of the most important factors is the colour of video frames which may vary a lot from frame to frame and from one camera to another. Thus, a colour normalization procedure must be carried out to reduce the illumination variations, and therefore to obtain invariant signatures. **Figure 4**a presents an example of a series of frames whose colours vary according to the illumination variations. Two series of silhouettes extracted from these frames are

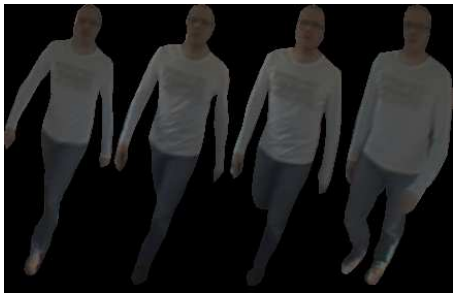
Taylor & Francis ©
Journal of Intelligent Transportation Systems: Technology, Planning, and Operations
http://www.tandf.co.uk/journals/titles/15472450.asp

presented: **Figure 4**b shows the series obtained without the colour normalization application, while figure 4c presents the results of the Greyworld normalization (Buchsbaum, 1980). We notice that the colours of silhouettes shown in **Figure 4**b (i.e. without invariant) change from frame to frame, meaning that the colour signatures extracted from these frames are not robust. Colour normalization reduces the illuminance variance (**Figure 4**c) and thus yields more robust colour signatures.

(a) Example of a series of frames captured inside the operating train



(b) Silhouettes extracted without colour normalization



(c) Silhouettes obtained using colour normalization.

Figure 4: Influence of illumination variations, and usefulness of colour normalization.

Hence, colour normalization must be applied to the silhouette of each person before computing its colour-based signature S. Many such methods have been proposed in the literature. In this paper, we only present the three invariances that lead to better results in our case:

- **Greyworld normalization** (Buchsbaum, 1980) 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:

$$I_k^* = \frac{I_k}{mean(I_k)} \tag{1}$$

  where I_k is the color value of channel k.

- **Histogram equalization** (Finlayson et al., 2005) 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 i and channel k is obtained with:

$$M_k(i) = \frac{\sum_{u=0}^{i} H_k(u)}{\sum_{u=0}^{Nb} H_k(u)} \tag{2}$$

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

Taylor & Francis ©
Journal of Intelligent Transportation Systems: Technology, Planning, and Operations
http://www.tandf.co.uk/journals/titles/15472450.asp

k .

- **Affine normalization** is defined by:

$$I_k^* = \frac{I_k - \mathrm{m}ean(I_k)}{\mathrm{s}td\left(I_k\right)} \tag{3}$$

The output of this first procedure, included in local image processors in normal mode, is the colour-position signatures, invariant to lighting conditions and estimated on each frame. For re-identification of a person between two cameras, the characterisation of each passage of a person is achieved by concatenating the signatures of several frames of the corresponding sequence. The final re-identification uses the algorithm presented in the following sections.

4 Dimensionality reduction for clustering and categorization

4.1 Overview

In statistics, dimensionality reduction is the process of reducing the number of random variables under consideration. It is an important procedure employed in various high-dimensional data analysis problems. It can be performed by keeping only the most important dimensions, i.e. the ones that hold the most information for the task, and/or by projecting some dimensions into others. A lower-dimensional representation of the data can be advantageous for further processing: classification, visualization, data compression, etc.

Over the past years, many dimensionality reduction techniques have been proposed. These techniques can be categorized into linear (e.g. Principal Components Analysis (PCA) (Hotelling, 1933) and Linear Discriminant Analysis (LDA, cf. Fischer, 1936)) and non-linear (e.g. Kernel PCA (Mika et al., 1999), Locally Linear Embedding (LLE) (Roweiss and Saul, 2000), Isomap (Tenenbaum et al., 2000), Laplacian Eigenmaps (Belkin and Niyogi, 2003), Hessian eigenmaps (Donoho and Grimes, 2003), Diffusion Maps (Nadler et al., 2005) and many variants of Spectral Analysis, such as Ng et al. (2001) or Von Luxburg (2007)) techniques. The differences between these two groups lie in their different assumptions and motivations: linear techniques assume that the data can be represented in a linear subspace of the high dimensional space, while non-linear techniques can manage complex non-linear data and attempt to preserve global and/or local properties of the original data in the lower dimensional representation. On very difficult databases, which is our case in this work, non-linear techniques may offer many more advantages compared to linear ones. In this article, we use dimensionality reduction to manage the signatures extracted from different image sequences corresponding to the passage of a person in front of several cameras. Two techniques have been compared on these datasets, a linear one (PCA) and a recent non-linear one (Laplacian Eigenmaps). In the following sections, after a few theoretical reminders of the two methods, we explain how we adapt them to our problem. This procedure helps us to compare the video sequences from the two cameras, and to make the final re-identification decision.

4.2 Principle component analysis approach

Principal Components Analysis (PCA) is the best linear dimension reduction

technique in the mean-square error sense. PCA constructs a low-dimensional representation of the data so that the greatest variance of the data comes to lie on the first few dimensions.

4.2.1 Mathematical formulation of PCA

Let $\mathbf{X}$ be a $M \times N$ matrix containing M vector $\mathbf{x}_i \in \mathrm{R}^N$ ($i = 1, \ldots, M$) stacked as rows in $\mathbf{X}$, and $\mathbf{W}$ be the zero mean normalized matrix deduced from $\mathbf{X}$ by subtracting each column of $\mathbf{X}$ from its mean and then normalizing by its standard deviation. Let $\mathbf{e} \in \mathrm{R}^N$ be a basic vector containing the coefficients of the linear combination in the new space. The projection of the corresponding data vectors from matrix $\mathbf{W}$ onto the basis vector $\mathbf{e}$ is the M dimensional vector $\mathbf{y} = \mathbf{We}$. The variance of $\mathbf{y}$ can be written as:

$$\frac{1}{M}\sum_{i=1}^{M} y_i^2 = \frac{1}{M}\mathbf{y}^T\mathbf{y} = \frac{1}{M}\mathbf{e}^T\mathbf{W}^T\mathbf{We} = \mathbf{e}^T\mathbf{Ce} \tag{4}$$

where $\mathbf{C} = \frac{1}{M}\mathbf{W}^T\mathbf{W}$ is the correlation matrix of $\mathbf{X}$.

The goal of PCA is to maximize the variance of $\mathbf{y}$. By imposing the constraint that $\mathbf{e}$ is a unity vector (i.e. $\mathbf{e}^T\mathbf{e} = 1$) and by applying the theory of Lagrange multipliers:

$$L(\mathbf{e}, \lambda) = \mathbf{e}^T\mathbf{Ce} - \lambda\left(\mathbf{e}^\mathbf{T}\mathbf{e} - \mathbf{1}\right) \tag{5}$$

we can find the critical points of the variance satisfying the constraint:

$$\frac{\partial L}{\partial \mathbf{e}} = \mathbf{Ce} - \lambda\mathbf{e} = \mathbf{0} \tag{6}$$

$$\Leftrightarrow \mathbf{Ce} = \lambda\mathbf{e} \tag{7}$$

Hence, the basic vector maximizing the variance of the projected data is given by the principal eigenvectors (i.e. principal components) of the correlation matrix of the zero mean data. The set of eigenvalues, ordered from large to small, is used to select a reasonable number of principal components $d << N$. The sum of the eigenvalues corresponds to the data variance, and a number of principal components can be selected so that the variance of the projected data reaches a reasonable percentage (e.g. 90%) of the initial variance.

4.2.2 Implementation of PCA for people re-identification

As mentioned in the introduction, the objective of this framework consists in re-identifying a person who has appeared in the field of one camera and then reappears in front of another camera. Let us suppose that the "re-identification CAM2-CAM1" function is turned on. k passages that have been recorded in the database of camera 1 are sent to camera 2. For each passage detected in front of camera 2, the signatures of the images of this sequence are extracted and used to perform re-identification.

Thus, a set of M signatures $S = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M\}$ consisting of the signatures belonging to k passages in front of camera 1 and the query passage in front of camera 2, is considered for the re-identification (a passage is characterized by the concatenation of the signatures of several images of the sequence denoted by n). Thus,

$M = n*(k+1)$. The input matrix $\mathbf{X}$ is a $M \times N$ matrix in which each row is a "colour-position" signature with $p = 200$. Such a matrix is illustrated in **Figure 5**.

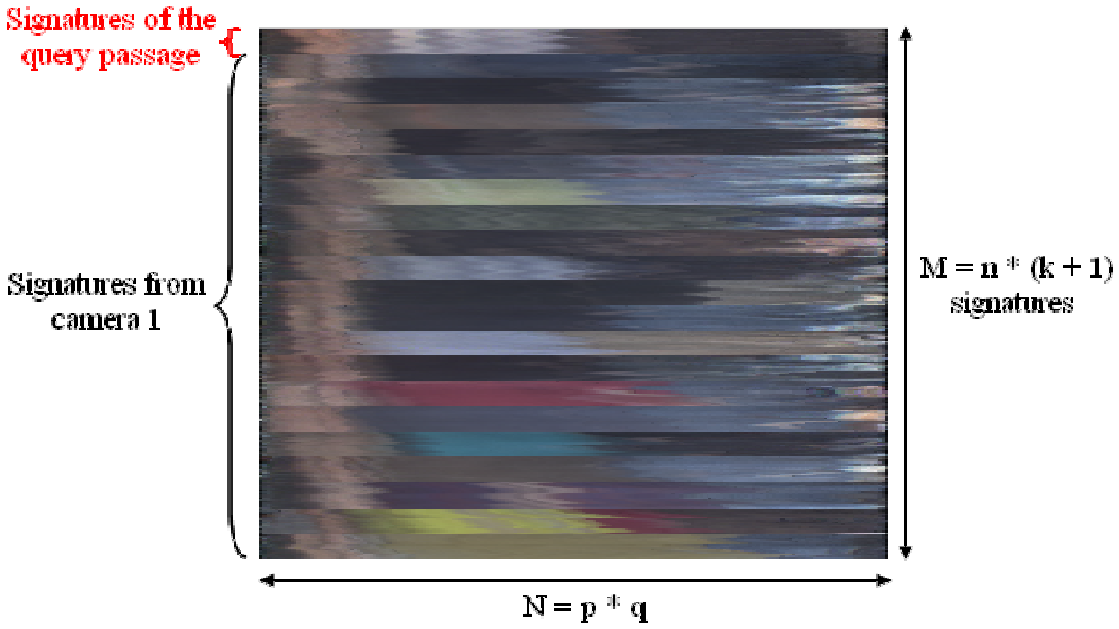


Figure 5: Example of a set of signatures.

By applying PCA matrix $\mathbf{X}$, we obtain the new coordinate system defined by the principal components (i.e. the eigenvectors sorted according to their eigenvalues). The set of signatures is projected into a lower dimensional space : the first 20 principal components corresponding to 96% of the total variance are kept. Since each passage is represented by the signatures of n frames, the barycentre of the n points obtained by projecting the n signatures into the new coordinate system is calculated. The distance between two barycentres is considered as the dissimilarity measure between two corresponding people. The larger the distance, the more dissimilar the two persons.

Hence, the dissimilarities between the query person, detected in front of camera 2, and each candidate person of camera 1, are calculated and then classified in increasing order. Then, in a perfect system, if the same person has been detected by the two cameras, the lowest distance between the barycentres should correspond to the correct re-identification. As it is not sure that the person walked under the two cameras, and as the measures are noisy in the databases used, we consider that all distances below a given threshold may correspond to a true re-identification.

4.3 Laplacian Eigenmaps approach

Laplacian Eigenmaps (LEM) (Belkin and Niyogi, 2003) is a technique for dimensionality reduction that preserves 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. The goal of LEM is to find embedding coordinates that minimize the sum of pairwise squared distances between the embedded points, weighted by the weight of edges in the neighbourhood graph. Using spectral graph theory, such a minimization is defined as an eigenproblem.

4.3.1 Principles and mathematical foundation

Let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\} \in R^N$ be M sample vectors. The LEM algorithm first constructs a neighbourhood graph $G = (V, E)$, where each data points $\mathbf{x}_i$ corresponds to a vertex v_i in this graph. Two vertices corresponding to two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by an edge that is weighted by $W_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{\sigma^2}\right)$. The first step of dimensionality reduction consists in searching for a new representation $\{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_M}\}$ with $\mathbf{y}_i \in R^M$ obtained by minimizing the cost function:

$$\phi_{LEM} = \sum_{ij} \left\| \mathbf{y}_i - \mathbf{y}_j \right\|_2 W_{ij} \tag{8}$$

Let $\mathbf{D}$ denote the diagonal matrix with elements $D_{ii} = \sum_j W_{ij}$ and $\mathbf{L}$ denote the un-normalized Laplacian defined by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The cost function can be reformulated as:

$$\phi_{LEM} = \sum_{ij} \left\| \mathbf{y}_i - \mathbf{y}_j \right\|_2 W_{ij} = 2\mathrm{Tr}\left(\mathbf{Y}^T \mathbf{L} \mathbf{Y}\right) \text{ with } \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M] \tag{9}$$

Hence, minimizing the cost function ϕ_{LEM} is proportional to minimizing $\mathrm{Tr}\left(\mathbf{Y}^T \mathbf{L} \mathbf{Y}\right)$. Dimensionality reduction is obtained by solving the generalized eigenvector problem

$$\mathbf{Ly} = \lambda \mathbf{Dy} \tag{10}$$

for the lowest non-zero d eigenvalues ($d << N$).

4.3.2 Implementation of Laplacian Eigenmaps for people re-identification

Similar to the PCA approach, the input of this method is also a set of signatures $S = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M\}$ consisting of M signatures belonging to k passages in front of camera 1 and one query passage in front of camera 2 (i.e. $(k+1)$ passages). We first associate to the set of signature vectors S a complete neighbourhood graph $G = (V, E)$ where each signature vector $\mathbf{S}_i$ corresponds to a vertex v_i in this graph. Two vertices corresponding to two vectors $\mathbf{S}_i$ and $\mathbf{S}_j$ are connected by an edge that is weighted by $W_{ij} = exp\left(-\frac{d\left(\mathbf{S}_i, \mathbf{S}_j\right)^2}{\sigma^2}\right)$. Here, we use the L^1 norm for computing the distance between two characteristic vectors $d\left(\mathbf{S}_i, \mathbf{S}_j\right) = \sum_{k=1}^{N} \left| \mathbf{S}_{ik} - \mathbf{S}_{jk} \right|$. The parameter σ is chosen as $\sigma = mean\left[d\left(\mathbf{S}_i, \mathbf{S}_j\right) \right]$, $\forall i, j = 1, \ldots, M$ $(i \neq j)$. Ideally, W_{ij} is large when two signature vectors indexed by i and j preferably belong to the same person, and is small otherwise. Now, we can compute the un-normalized Laplacian $\mathbf{L}$ and produce the eigenstructure of $\mathbf{L}$. The eigenvectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ provide a new coordinate for the image set. Dimensionality reduction is obtained by considering the lowest eigenvectors with $d = N$. In our current problem, we use the 20 lowest eigenvectors to create the

new coordinate system. The dimensionality reduction operator can be defined as $h:\mathbf{S}_i \rightarrow \mathbf{u}_i = [y_1(i),\ldots,y_{20}(i)]$ where $y_k(i)$ is the i^{th} coordinate of eigenvector $\mathbf{y}_k$.

The following steps are carried out in the same manner as described in the PCA-based approach. Thus, for each query passage, k distances are calculated and the candidate people corresponding to the distances below a decision threshold are chosen as the result of re-identification.

5 Experiment and results

In the previous section, we recalled the theoretical properties of two dimensionality reduction techniques and described how to adapt them to our problem of re-identification. In this section, we perform the evaluation of our proposed algorithms by using two real datasets which are described in the subsection 5.1. The setup of our experimentations and the obtained re-identification results are presented in subsection 5.2.

5.1 Datasets for the evaluation

As mentioned above, our research aims to set up an onboard surveillance system able to re-identify a person through multiple cameras. Before collecting the onboard dataset, another dataset acquired in INRETS premises was collected for the evaluation of our algorithms. In the framework of the European BOSS project, a system composed of two cameras was installed onboard a train to collect the second dataset. These two datasets are described below.

5.1.1 First dataset

The first dataset, collected on INRETS premises, contains video sequences of 40 people captured by two cameras. We have chosen two different locations: indoors in a hall lit by windows, and outdoors with natural light. **Figure 6** illustrates one of the forty people in these two different environments.

Figure 6: Illustrations of the first database representing the same person in two different environments: indoors in a hall (left) and outdoors (right).

In order to characterize the passage of an individual, and for processing time purposes, we do not take into account all the frames of a video sequence, but only a subset of regularly spaced frames covering the whole passage. For this first dataset, we have chosen to extract ten frames per person and per location. This number of frames is

sufficient for describing the characteristics of a passage while ensuring real-time processing. **Figure 7** illustrates, for a given person, the extracted silhouettes for the selected frames in the two locations.

Figure 7: Example of frame extractions for two sequences of the same person in two different locations: indoors in a hall (top); outdoors in a garden (bottom).

5.1.2 Train dataset

The second dataset is collected by a system composed of two cameras installed onboard a train at two different locations: one in the corridor and one in the cabin. This dataset contains video sequences of 35 people, each acquired by both cameras. This dataset is more difficult than the first one, since these two cameras are set up with different angles and the image quality is impaired by many factors, such as fast illumination variations, reflections, vibrations. **Figure 8** illustrates this data set.

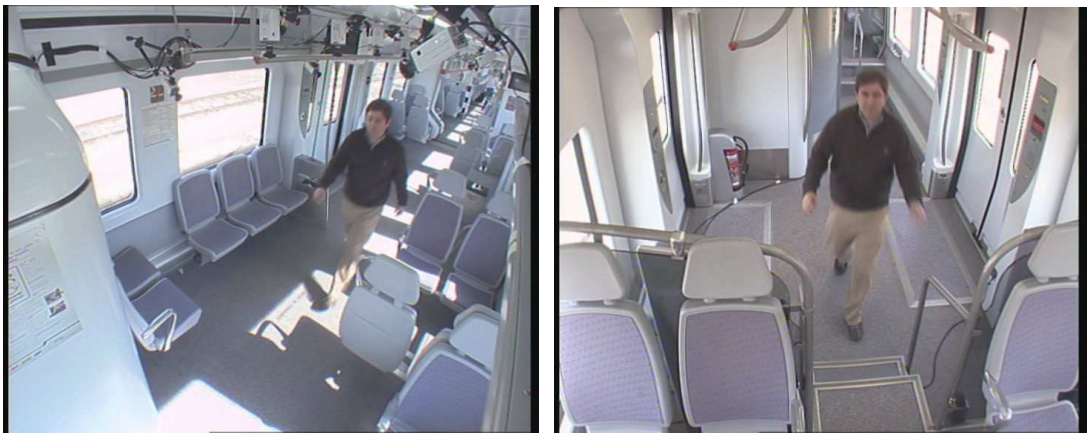


Figure 8: Illustrations of the second database representing the same person in two different environments: in the cabin (left) and in the corridor (right).

Due to the difficulties of this dataset, we extract 20 frames for each passage.

5.2 Experimental results

As described in section 4, for each query passage in front of one camera, the distances (i.e dissimilarities) between the query person and each of the candidate people

of the other camera are calculated. A decision threshold is chosen; distances below the threshold indicate a re-identification (score = 1), meaning that the set of signatures represents the same person. Distances above the threshold mean distinction: the test signature set represents different people (score = 0).

Let K be the number of people for each location. $K \times K$ distances are calculated and then compared with the threshold. The resulting scores can be arranged in a $K \times K$ 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). An actual re-identification system can yield one of four possible results:

- True re-identification (true match, true positive): the system declares a re-identification (score = 1) when the two passages belong to the same person (the diagonal elements).
- True distinction (true non-match, true negative): the system desclares a distinction (score = 0) when the two passages represent two different people (the off-diagonal elements).
- False re-identification (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 elements).
- False distinction (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 elements).

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 refering to the Equal Error Rate (EER) point at which the FRR is equal to the FDR.

In order to illustrate the performances of our system, we use the ROC curve (Quddus et al., 2006) which is a plot of true re-identification rate (TRR) versus false re-identification rate (FRR) as the threshold value varies. These two rates can be calculated from the score matrix using the following definitions:

$$\mathrm{TRR} = \frac{\sum_{k=1}^{K} \left(score_{kk} = 1 \right)}{K} \tag{11}$$

$$\mathrm{FRR} = \frac{\sum_{k=1}^{K} \sum_{l=1}^{K} \left(score_{kl} = 1, k \neq l \right)}{\mathrm{K}\left(K - 1 \right)} \tag{12}$$

In **Figure 9**, divided into four parts according to two function modes (CAM2-CAM1 or CAM1-CAM2) and two approaches (PCA approach for the first two diagrams and LEM approach for the following two diagrams), we find the ROC curves obtained for the first database. The ROC curves highlight the trade-off between the TRR and the FRR: an increase in TRR is accompanied by an increase in FRR. The closer the ROC curve approaches the top left-hand corner of the plot, the better the method is. The optimal points (i.e. the points that assume the equality of FRR and FDR - EER points)

are also presented by the crossing between the ROC curves and the EER.

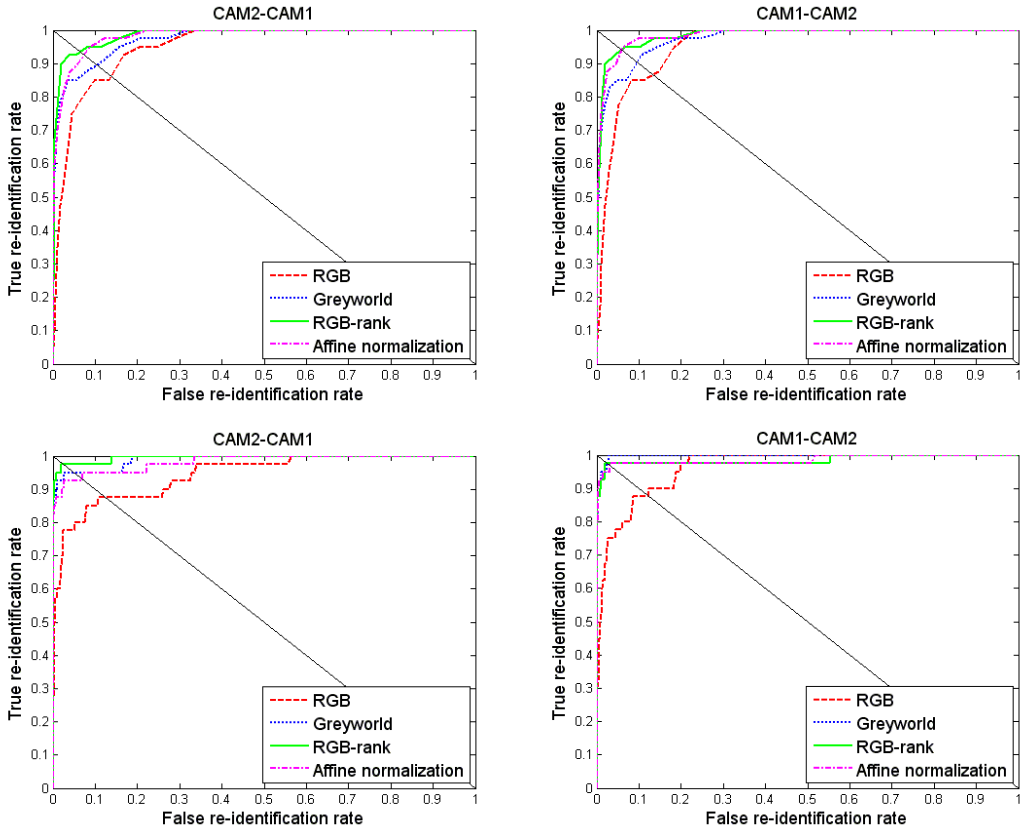


Figure 9: ROCs obtained by testing the first database (top: results of PCA approach; bottom: results of LEM approach).

For the first database, we obtain very satisfying optimal rates for both approaches. The LEM approach leads to slightly better results. The invariant normalisations actually improve the results compared to the RGB space. Table 1 summarises the comparative results of RGB space and three normalization procedures for the LEM approach.

	CAM2-CAM1	CAM1-CAM2
RGB	88%	88%
Greyworld	95%	98%
RGB-rank	98%	98%
Affine normalization	93%	98%

Table 1: TRRs at the optimal points corresponding to RGB space and three normalization procedures obtained using the LEM-based approach.

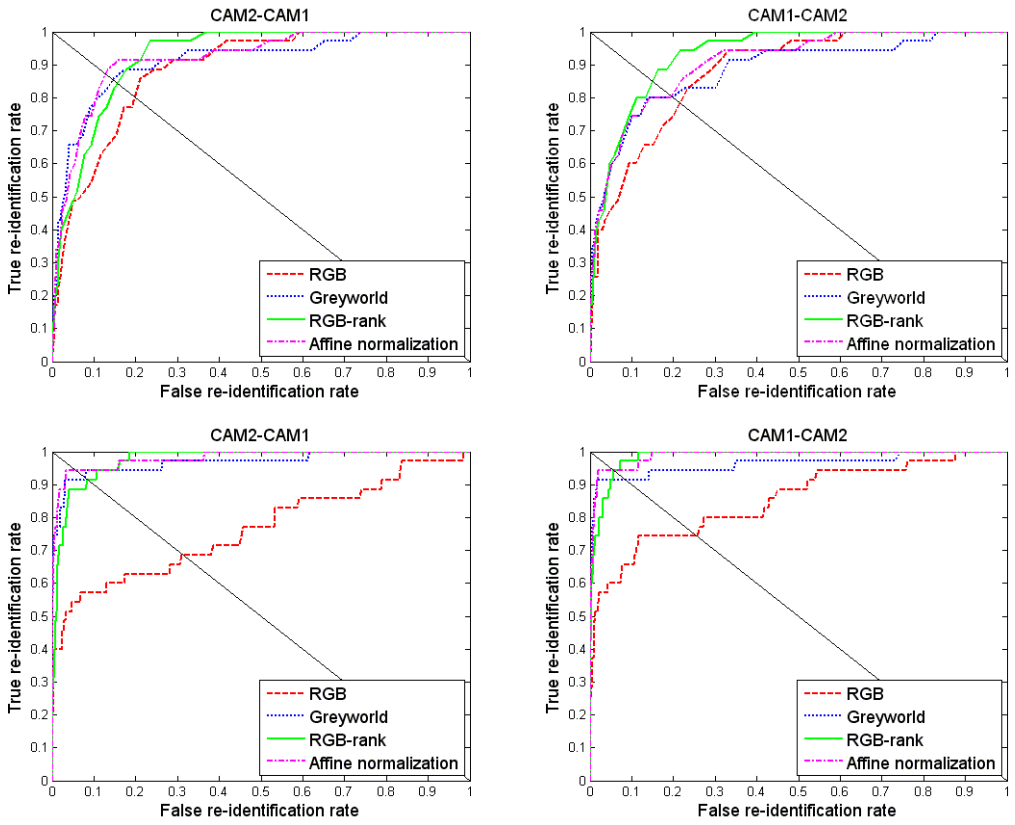


Figure 10: ROC curves obtained by testing the second database (top: results of the PCA approach; bottom: results of the LEM approach).

For the second, much more difficult database, the same kind of results are presented in **Figure 10**. In spite of the difficulties encountered in the processing of this database (moving train, illumination changes, sun not shining regularly on the windows, many shadows), the global performances of the two approaches are still very satisfying (87% for PCA, 94% for LEM). Nonetheless, the LEM approach yields better results than the PCA one. Once again, using the illumination invariants improves the results upon the RGB space (see **Table 2** for the comparative results of the LEM approach).

	CAM2-CAM1	CAM1-CAM2
RGB	68%	74%
Greyworld	91%	91%
RGB-rank	91%	94%
Affine normalization	94%	94%

Table 2: TRRs at the optimal points corresponding to RGB space and three normalization procedures obtained by using the LEM-based approach.

Hence, the combination of the colour position signature with a non-linear clustering method and an illumination invariant procedure allows us to achieve very good performances in the difficult task of video comparison. Graph-based methods in general, and LEM method in particular, outperform the classical dimensionality reduction techniques such as PCA and LDA, since can learn the global structure of the

manifold and preserve the properties of the original data in the low-dimensional representation. The good performance of the LEM method with our very difficult dataset encourages our future work aiming to improve the re-identification results by using the non-linear graph-based approach.

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 the problem of re-identification without using the decision threshold is based on the nearest-neighbor algorithm. The distances between the query passage and all the candidate passages captured by another camera are classified in increasing order. The closest candidate 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. By using this method of evaluation, we obtain similar results to those previously obtained.

5.3 Influence of the number of extracted frames on the global performance of the system

In this section, we provide some results on the influence of the number of frames used on the global performances of the system. This survey is carried out only on the second dataset (onboard a train data) and with the LEM-based approach. The best number of frames to use could be delicate to determine, because it must represent a compromise between the good acuracy of the system and an acceptable processing time.

Table 3 provides some performance results in terms of true re-identifaction rates when we cross the number of frames used with the nature of invariance method. In the table, the results concerning the RGB space are omitted as, given the nature of the onboard data and the difficulties with regard to variations in lighting, it yields the worse results. As a result, **Table 3** concerns the three invariant methods tested so far.

True re-identification rate	Greyworld normalization	RGB-rank normalization	Afffine normalization
5 frames	86%	86%	89%
10 frames	89%	89%	91%
15 frames	89%	91%	91%
20 frames	91%	91%	94%

Table 3: Influence of the number of frames on the true re-identification accuracy (on the onboard train dataset and with LEM approach).

From the table, we can note that the higher the number of frames used, the higher the rate of the re-identification, which was to be expected. On the other hand, with the accuracy already achieved with 20 frames (91% for Greyworld, 91% for RGB-rank and 94% for affine normalization), we can easily imagine that the accuracy will not be much more improved with additional frames. This means that 20 frames are sufficient to obtain satisfying rates of re-identification. The choice of 20 frames within a sequence seems to be a good compromise.

We will see in the following paragraph that, with 20 frames, it is possible to achieve

a real-time implementation while maintaining a highly satisfactory rate of re-identification.

Our approach yields a better global re-identification rate than similar works in the literature, especially considering the large and diverse dataset used. For instance, Gheissari et al. (2006) used a dataset with 44 people and obtained a 60% true re-identification rate. Wang et al. (2007) achieve a matching rate of 82% on a dataset of 99 persons. In the system proposed by Yu et al. (2007), the accuracy is 95% for a 30 people dataset captured indoors. Nakajima et al (2003) obtained a 100% accuracy but on a very reduced dataset (4 people).

5.4 Real-time considerations

For the moment, the whole final system composed of two cameras and the supervisor computer are not permanently installed in the train for evaluation purposes. Therefore, the figures we provide here on the processing time concern an on-line simulation of the network in our laboratory with data coming from the train.

Since our research aims to establish a real-time surveillance system, the computational complexity and the processing time of the algorithms are critical. Here, we provide the computational complexity of the PCA and LEM approaches as well as the processing time of the experimentations carried out on a PC computer with Intel Core 2 Duo 2.1GHz CPU. This is done for the second dataset.

The whole system is composed of three main functionalities :

- Raw-frame processing for the foreground extraction (extraction of silhouettes),
- Colour-position signatures calculation
- Re-identification, using either the PCA or the LEM approach.

For silhouette extraction, we use the background subtraction algorithm combined with the morphological operators and the CCL algorithm, which take 0.05s for processing one frame. For the colour-position signature calculation, we set the total number of equal parts of the silhouette to $p = 200$. Signature extraction takes 0.1s for one frame. This ensures real-time processing to characterize a passage with $n = 20$ extracted frames.

For the re-identification part, the computational complexity and the processing time of the two algorithms, PCA and LEM, are determined by the total number of frames $M = n*(k+1)$ and the original dimensionality is $N = p \times q$. For the second dataset, the total number of frames for one query is $M = 20*(35+1) = 720$ frames and the dimensionality is $N = 200 \times 3 = 600$.

For the PCA-based approach, the computation of the correlation matrix has a computational complexity of $O(MN)$ and the eigenanalysis of the $N \times N$ correlation matrix is performed in $O(N^3)$. Thus, PCA-based algorithm has a total computational complexity of $O(MN) + O(N^3) = O(720 \times 600) + O(600^3)$. The global processing time for one query with the PCA-based algorithm is therefore 2.2 seconds.

For the LEM-based approach, the Gaussian kernel computation and the eigenanalysis of an $M \times M$ matrix are performed in $O(NM^2)$ and $O(M^3)$ respectively. Thus, LEM-

based algorithm has a total computational complexity of $O(NM^2)+O(M^3)=O(600\times720^2)+O(720^3)$. The processing time for one query is 6.7 seconds.

We notice that the LEM-based approach requires more computational resources than the PCA-based approach. The processing time of 6.7s of the LEM-based algorithm, which is the better method for re-identification, is fairly high for a real-time implementation. However, this is not a real problem because several improvements can be carried out:

- Optimization of the software
- More powerful computer utilization and specifically parallel architectures
- Lower frame-resolution (current size : 720*576 pixels)

With these three improvements, the overall processing time could be divided at least by 5. Laboratory tests are in progress.

6 Conclusion and perspectives

In this paper, we have presented the architecture of a smart surveillance system including multiple sensors and a specific algorithm to re-identify people who appear and reappear at different times, across different cameras. Our approach relies on the extraction of an appearance-based signature from each frame of the sequence characterizing the passage of a person, combined with an algorithm that reduces the effective working space and realizes the comparison of the set of signatures. The first step of the system thus consists in estimating a feature that describes the person in each frame of the sequence. It must preserve the useful information, be discriminating enough to separate different people, and be unifying enough to make coherent classes with all the features belonging to the same person. The proposed solution, called ”colour-position” signature, combines spatial and colour information and thus keeps a large amount of information. The unifying power of the feature is reached thanks to illuminant invariant methods, able to describe a person regardless of the lighting. The passage of a person is then obtained by characterising several frames. This leads to a large quantity of data, that must be reduced and processed. To do this, two dimensionality reduction techniques are tested. The first one, which is linear, is the classical Principal Components Analysis, while the second one is a recent non-linear method called Laplacian Eigenmaps. Once the reduction is achieved, a distance is introduced to determine whether two sequences depict the same person or not.

The global system was tested on two real data sets. The first one was acquired in laboratory, while the second comes from a real situation. Specifically, a surveillance system including two sensors was installed on board a train and a realistic data set was collected. The experimental results show that the non-linear approach, used for dimensionality reduction, always yields the best results: 98% true re-identification rate for the first database and 94% for the second one. These results, stemming from the combination of the colour-position signature, the dimensionality reduction techniques and the illumination invariance, are very satisfactory.

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 is hard to carry out. Thus, in order to make available to researchers a

common, extensive database for pertinent comparison, the dataset shot in real-life conditions onboard a moving train in the framework of the BOSS project can be found on the website http://www.multitel.be/boss. It comprises the challenges due to fast illumination variations, reflections, vibrations, and static/dynamic occlusions that perturb the current video interpretation tools.

Several perspectives are envisaged to improve the performance of our system, such as the introduction of temporal information (e.g. the travel time from one camera to another) or moving direction of people. A fusion of global and local descriptors (e.g. descriptors using interest points) is also considerated. We also envisage, in the medium term, to add a complementary step of tracking within a single camera in order to handle more challenging scenarios (multiple passages in front of cameras, occlusion, partial detection,…).

References

Ballard, D. H. and Brown C. M. (1982). Computer Vision. Upper Saddle River, NJ: Prentice Hall Professional Technical Reference.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15(6)**,1373-1396.

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

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

Donoho, D. L., and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, **100(10),**5591-5596.

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

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**,179-188.

Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, 1528-1535.

Harville, M. (2002). A framework for high-level feedback to adaptive, per-pixel, mixture-of gaussian background models. *Lecture Notes in Computer Science*, 543-560.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441.

Javed, O., Rasheed, Z., Shafique, K., and Shah, M. (2003). Tracking across multiple cameras with disjoint views. Ninth IEEE International Conference on Computer Vision.

Kim, K., Chalidabhongse, T. H, Harwood, S., and Davis, L. (2004). Background

modeling and substraction by codebook construction. *International Conference on Image Processing, ICIP'04*, **5**.

Kohonen, T. (1998). Learning vector quantization.

Mika, S., Scholkopf, B., Smola, A., Muller, K. R., Scholz, M., and Ratsch, G. (1999). Kernel pca and de-noising in feature spaces. *Advances in Neural Information Processing Systems*, **11(1)**, 536-542.

Nadler, B., Lafon, S., Coifman, R., and Kevrekidis I. (2005). Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *Advances in Neural Information Processing Systems*, 955-962.

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

Ng, A. Y., Jordan, M., I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, **2**, 849-856.

Quddus, M. A., Ochieng, W. Y. and Noland, R. B. (2006). Integrity of map-matching algorithms. *Transportation Research Part C: Emerging Technologies*, **14(4)**, 283-302.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290(5500)**, 2323-2326.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**, 246-252.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290(5500)**, 2319-2323.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17(4)**, 395-416.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P. (2007). Shape and appearance context modeling. *IEEE 11th International Conference on Computer Vision*, 1-8.

Yoon, K., Harwood, D., and Davis, L. S (2006). *Journal of Visual Communication and Image Representation*, **17(3)**, 605-622.

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