Oliver Moll, J.; Albiol Colomer, A.; Albiol Colomer, AJ.; Mossi García, JM. (2016). Using latent features for short-term person re-identification with RGB-D cameras. Pattern Analysis and Applications. 19(2):549-561. doi:10.1007/s10044-015-0489-8.

Pattern Analysis and Applications manuscript No.
(will be inserted by the editor)

Using Latent Features for Short-Term Person Re-Identification with RGB-D Cameras

Received: date / Accepted: date

Abstract This paper presents a system for people re-identification in uncontrolled scenarios using RGB-depth cameras. Compared to conventional RGB cameras, the use of depth information greatly simplifies the tasks of segmentation and tracking. In a previous work, we proposed a similar architecture where people were characterized using color-based descriptors that we named bodyprints. In this work, we propose the use of *latent feature* models to extract more relevant information from the bodyprint descriptors by reducing their dimensionality. *Latent features* can also cope with missing data in case of occlusions. Different probabilistic *latent feature* models, such as Probabilistic Principal Component Analysis and Factor Analysis, are compared in the paper. The main difference between the models is how the observation noise is handled in each case. Re-identification experiments have been conducted in a real store where people behaved naturally. The results show that the use of the *latent features* significantly improves the re-identification rates compared to state-of-the-art works.

Keywords Bodyprint · Probabilistic PCA · Factor Analysis · Missing Data · Re-Identification · Surveillance · Person Detection · Appearance Matching · Kinect

1 Introduction

Vision-based systems have become essential in almost every business. The deployment of camera networks has become widespread during recent years in domains such as surveillance, marketing, and sports.

The ubiquity of these vision systems, in many cases, allows statistical information about individuals such as trajectory, velocity, behaviour, and occupancy to be obtained. In order to extract such information, a person's identity needs to be tracked through time and space.

For camera networks, an association mechanism is also required to track people across different cameras. In the case of overlapping cameras, association turns out to be trivial since a space-time constraint can be applied in calibrated scenarios. However, in a general situation with non-overlapping cameras, the association is generally handled by re-identification [11]. Re-identification is still an open problem since many difficulties may arise in unconstrained scenarios, such as: changes in illumination, different points of view, pose variations, occlusions, and behavioural differences.

The recent appearance of RGB-Depth cameras, such as Microsoft Kinect [1], offers the opportunity to explore the use of many emergent three-dimensional computer vision techniques in surveillance scenarios at an affordable cost. In this work, we make use of kinect cameras for tracking and describing individuals. A person's appearance is described in a first stage by dividing the person into horizontal strips and extracting the temporal mean colors for each strip. The main contribution of this work is the introduction of the *latent features* that are generated using probabilistic latent variable models applied over the appearance features. The use of *latent features* circumvents some problems that appear when directly using appearance features such as missing data, acquisition noise, outliers, and high correlation of color features. Several metrics are also proposed to match individuals in the re-identification task.

The remainder of the paper is organized as follows. Section 2 reviews related work in the area of person re-identification. In Section 3 , a general overview of the system is presented, and the main concepts on *latent features* and the equations used to extract them from observed features are provided in Section 4. Different matching metrics, which are used to compare *latent features*, are described in Section 5. The performance of the system is evaluated in Section 6 in

a real scenario. Finally, some conclusions are drawn in Section 7.

2 Background

This section presents a review of the most relevant methods for person re-identification. Additionally, a short introduction of dimensionality reduction techniques based on probabilistic latent variable models is given in Subsection 2.2.

2.1 Techniques for person re-identification

A common assumption in the literature, which we also consider in this work, is that there is no change of clothes for each person in all the views. With this assumption, the re-identification problem focuses on how to describe the appearance of individuals and how to match them among cameras. A description of techniques grouped by affinity is presented below.

Appearance is usually described using features of color, texture and shape. However, color and texture techniques are by far the most widely used in the literature. Appearance methods for person re-identification are also commonly grouped into single or multiple-shot methods [11]. Single-shot methods use only one image to perform identification, while multiple-shot methods use different images of the same person obtained by tracking. Multiple-shot methods can exploit other contextual cues such as spatio-temporal reasoning [21].

Color histograms, which are easily extracted and scale invariant, have been proposed in many works [25] [27] [34] [18] [36] [29] [49]. Bazzani et al. [11] use color global histograms combined with recurrent local patterns that characterize texture after epitomic analysis. The main problem associated with color histograms is that spatial information is discarded. In order to avoid this problem, Dikmen et al. [16] divide the image into a grid of fixed cells where color histograms are computed at each cell. Other authors [12] divide the image into horizontal strips and characterise color features for each strip. Bak et al [6] use Mean Riemannian covariance patches for describing feature distributions, considering temporal information of appearance. Mazzon et al [33] use a centered patch in the upper body to describe the color of a person's appearance. They complement appearance information with contextual data in order to filter people according to potential paths that they can follow.

Many different color spaces are proposed in the literature to represent color information. However, the most common choices are RGB [36] [27] [15] [11] and HSV [34] [23] [18]. A common problem found in re-identification is the variation of illumination conditions among cameras. For this reason, some authors [32] propose algorithms to compensate for these variations.

Texture information is usually retrieved using well-known local descriptors such as SURF [34], SIFT [28] [43], and Haar [4]. Farenzena et al. [18] use texture information by searching recurrent local motifs with high entropy. In many, cases texture information is concatenated with color information into a long descriptor as in [40], where LBP features are merged with RGB and HSV color histograms. Another alternative for fusing color and texture information is proposed in [30], where LBP features based on the quaternionic representation are used.

Re-identification algorithms can also be grouped into holistic [17] and part-based [5] [45]. Although part-based methods are very promising, holistic methods are still more robust in challenging scenarios [17]. Some other authors use other passive biometrics such as face [10], gait [46], and iris [42]. However, the low resolution of the images, occlusions, and the variety of poses make biometric-based techniques ineffective in many scenarios.

Many of the existing re-identification methods share the idea of representing the human body as a bag of instances described by appearance descriptors. However, an alternative matching framework is presented in [39]. In this framework, individuals are represented by means of a vector of dissimilarity values with a set of stored prototypes. In [22], a model composed of different orientations of the person is used. Each orientation is estimated from the person's trajectory and modelled by a different feature vector.

Recently, several authors have been introducing 3D models to describe a person's appearance. Baltieri et al [7] propose a 3D generic rigid body model that is filled up using person's appearance acquired with 2D calibrated cameras. Papadakis et al. [35] introduce a cylindrical 3D descriptor for generic object re-identification in controlled scenarios. The cylinder is filled up with the projection of a set of 2d panoramic views of the object.

The use of RGB-depth cameras has recently been proposed for person re-identification [3]. The use of these sensors eases the tasks of person segmentation and tracking and allows calibrated virtual views of the person to be created. On other hand, Barbosa et al. [8] use Kinect cameras to extract 3D soft biometric cues such as skeleton and surface-based features that are invariant to appearance variations. Their approach is indicated when handling long-term re-identification problems where a person's appearance may change over time. However, this information does not provide reliable results by itself.

2.2 Learning methods for re-identification

The paradigm of person re-identification can be commonly addressed as a learning-based problem given the set of fea-

tures that describe a person in the dataset. Several works such as [25] [20] [49] use supervised learning to perform feature selection. As a representative work, Zheng et al. [49] formulate this problem as a relative distance comparison learning problem in order to find the optimal similarity measure for several images of the same person. Other researchers use unsupervised reduction techniques such as Dictionary Learning [48], Manifold Learning [31] or PCA-LFDA [38] to find discriminant features.

In this paper, we follow an unsupervised learning approach by using probabilistic dimensionality reduction techniques. Dimensionality reduction techniques have been widely used in machine learning for the purposes of data visualization, data compression, noise removal, pattern recognition, exploratory analysis, and time series prediction. Depending on the nature of the observations, techniques can be classified [19] into linear methods such as Principal Component Analysis (PCA), Factor Analysis (FA) or Projection Pursuit (PP), and non-linear methods such as Independent Component Analysis (ICA) or non-linear PCA, just to cite a few.

PCA is the most extended method. It does a linear projection of the observed data onto a subspace of lower dimensionality such that the variance of the projected data is maximized. A probabilistic treatment of PCA (PPCA), which can be expressed as a probabilistic latent variable model problem, has been proposed in [41] and [37]. PPCA is based on a linear-Gaussian framework in which all of the marginal and conditional distributions are Gaussian. The probabilistic treatment has the advantage of elegantly solving some problems of direct PCA such as missing data due to occlusions and outliers.

Other probabilistic methods, such as Factor Analysis (FA) [9], are also linear Gaussian latent variable models that are similar to PPCA. The only difference between them relies on the covariance matrix of the observed data given latent variables. In PPCA, it is full covariance matrix, whereas in FA it is diagonal. In this paper, we will explore the use of different probabilistic models to extract *latent features*, as explained in Section 4.2.

3 System Overview

The re-identification system proposed in this paper is intended to work in uncontrolled environments where people can move and behave freely. An example of such a scenario is a store where two cameras cover the entry and exit areas, respectively. In this scenario, we want to re-identify the people that entered the store at the exit. The cameras are assumed to cover non-overlapping areas and the illumination conditions between cameras can be different. Althoughwe only consider the case of two cameras in this set up, the system can be extended to work with more cameras.

Figure 1 shows the block diagram of the system for one camera node. After scene calibration [3], given the depth of a pixel (provided by the sensor), it is possible to obtain its height with respect to the ground. Knowing the height of every pixel allows a *Height Map* to be built [47]. A Height Map looks like a top-view of the scene where the pixel values represent the height of the highest point of the person in the image captured by the sensor as a function of the ground coordinates. Figure 1 shows an example where the depth image is converted to a Height Map. People can be easily detected as local maxima of the Height Map as shown by the red circle in Fig. 1.

The tracking module receives detections of people and either assigns them to pre-existing tracks or starts new tracks. It is possible to track several people simultaneously. The tracker allows to gather information of the same person over time.

For each frame of a tracked person, we create a vector with the appearance information. Each element of this vector contains the mean color of a horizontal strip at each height. These vectors are appended over time, creating a temporal color signature for a person (Fig. 1). The temporal signature is finally summarized by a *bodyprint*, which contains the mean color together with its variance for each height (Fig. 2 shows a few examples of bodyprints). The bodyprint extraction also entails basic color normalization to deal with changes in illumination and cameras. For a more detailed explanation of this step, we encourage the reader to consult our previous work [3].

Finally, probabilistic latent variable models are used to extract *latent features* from the bodyprints. These features provide a solution to cope with noisy data, outliers, and occlusions, which generate missing data in the descriptor grid. Re-identification is performed by comparing *latent features* of people captured from different cameras. The next section provides detailed information about the *latent feature* extraction step.

4 *Latent Feature* Extraction

In this section, we first discuss some of the weaknesses of bodyprints. To get around these difficulties, different *latent feature* models are proposed in Section 4.2. In the sequel, matrices are represented in bold upper case, column vectors are represented in bold lower case and real scalars are represented in italic lower case.

4.1 Motivation

Bodyprints were introduced in our previous work [3]. Although this descriptor achieves remarkable re-identification

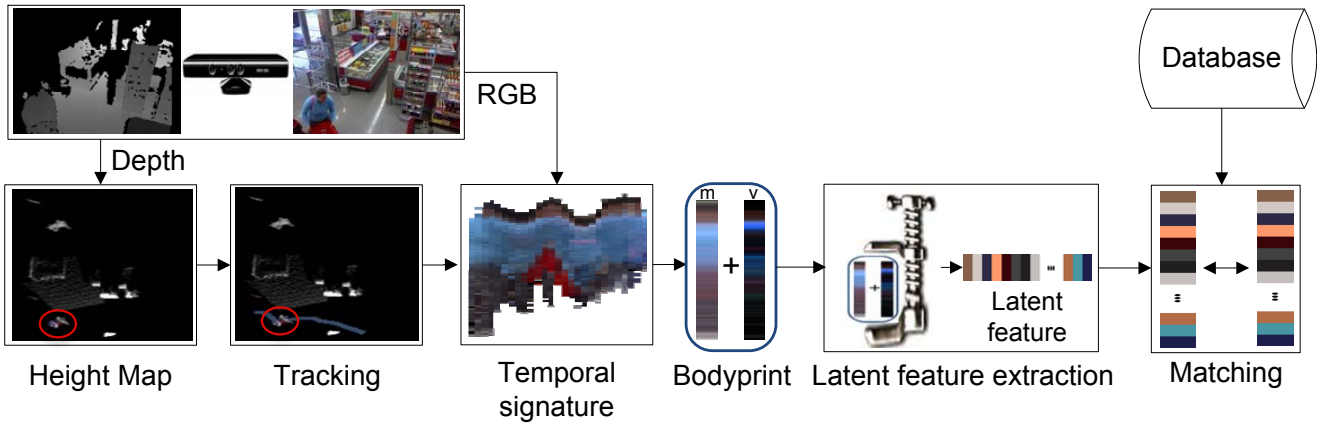

Fig. 1 Latent feature extraction for a single node in the camera network. The detected and tracked person is highlighted using a red circle in the Height Map and Tracking modules.

Fig. 2 Some examples of bodyprints. All the examples show different outliers generated by carried objects. Figure c presents missing data in the lower body.

rates compared to other state-of-the-art methods, its performance may be degraded in complex scenarios by some environmental variables, such as non-uniformly distributed changes in illumination, the presence of carried objects, occlusions, etc. Table 1 lists several of these extrinsic factors and how they affect the bodyprints. In addition, other intrinsic parameters (related to the descriptor), such as the height of the horizontal strips can also affect performance since thin strips generate high dimensional and correlated feature vectors.

Figure 2 shows some challenging examples. A representative frame of the tracked person is shown next to the temporal signature and its corresponding bodyprint in each example. The first column of the bodyprint represents the temporal mean color at each height and the second column represents the variance of the three color channels. In Fig. 2 (a), a person enters the store and gets a red shopping basket from a stack of baskets. In the temporal signature, the shopping basket appears as an outlier that increases the red color variance especially in the lower part, as shown in Fig. 3. Other carried objects, such as the child shown in Fig. 2 (b) or the shop items shown in Fig. 2 (c) and Fig. 2 (d), also produce outliers in the bodyprint and increase the color vari-

Table 1 Examples of surrounding and enviromental factors and how they affect the bodyprints.

Extrinsic factors	How the features are affected
Non-uniform changes in illumination	Non-uniform color variance
Pose variations	Non-uniform color variance
Gait	Non-uniform color variance
Occlussions	Missing data
Carried objects	Outliers or non-uniform color variance
Flat-colored clothes	Highly correlated features

ance at the corresponding heights. Note that they should be treated as outliers because the person's appearance is different from entrance to exit. Finally, in Fig. 2 (c), the feet of the person are never visible, so the corresponding parts of the bodyprint are missing.

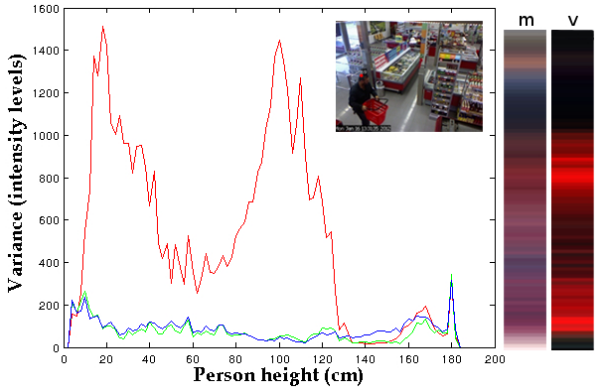

Fig. 3 Color variance of a contaminated bodyprint. The appearance of the red shopping basket produces a high variance in the red channel.

4.2 Latent feature models

The previous examples show that bodyprint features are highly correlated. This fact suggests that bodyprints lie in a much lower dimensional space, and, hence, the use of dimensionality reduction techniques seems appropriate to extract relevant features. An additional advantage of reducing the dimensionality of the descriptor is that noise that does not lie in the reduced feature space can be removed, thereby alleviating the problems caused by outliers.

Classical dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) cannot cope with missing data and they do not consider additional information such as the variance of the observed features. For this reason, we propose applying probabilistic latent variable models that are equivalent to classical techniques in the case of non-missing data and also allow information about the variance of the observed data to be incorporated. Note that the variance contains information about the confidence of observed feature values.

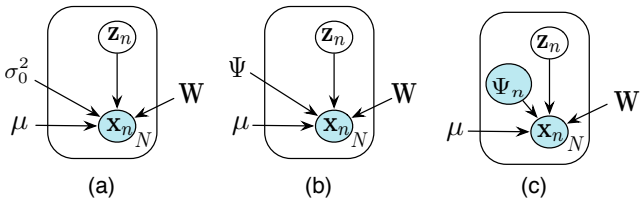

Fig. 4 Different latent-feature extraction models. In (a), the noise is considered to be isotropic and constant for all samples. In (b), the noise is modeled by a diagonal matrix and is constant for all samples. In (c), there is a different diagonal matrix representing the noise for each sample.

Probabilistic *latent feature* models are usually represented by graphs [13], as shown in Fig. 4. In the graph, model variables are placed in circles. The difference between white circles and blue circles is that in the blue circles, the variables are observed or measured, whereas in the red circles, the variables are not observed and can be inferred from observed features. This is why they are called *latent features*. The rectangular box surrounding the circular variables in the graph represents a plate. Each plate embeds N training samples of which only a particular $\{\mathbf{x_n}, \mathbf{z_n}\}$ pair is depicted. Both $\mathbf{x_n}$ and $\mathbf{z_n}$ are vectors of length l_x and l_z, respectively. While $\mathbf{x_n}$ represents the observed color features of a person (bodyprints), $\mathbf{z_n}$ denotes the corresponding latent variables ($l_z < l_x$) that we want to extract by inference. Terms outside the plate indicate the model parameters: mean, projection matrix and noise that are shared among all training and test samples. These parameters are obtained during the training stage using the EM algorithm [37] because there is no a maximum likelihood closed solution for some of the proposed models .

In the following, we introduce the different probabilistic *latent feature* models that we have used in this work. The main difference among models is how noise is handled in each case.

Probabilistic PCA (PPCA). The purpose of PPCA is to capture the covariance structure of an observed dataset by assuming a linear transformation between the latent and observed spaces. In PPCA, all marginal and conditional distributions are assumed to be Gaussian. The equations of the

generative model are:

$$\mathbf{x_n} = \mathbf{W}\mathbf{z_n} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \quad (1)$$

where $\mathbf{W}$ is an $l_x \times l_z$ linear transformation matrix that converts from latent to observed spaces, $\boldsymbol{\mu}$ is an l_x vector that represents the model mean, and $\boldsymbol{\epsilon}$ is an l_x zero-mean isotropic Gaussian noise vector, $p(\boldsymbol{\epsilon}) = N(\boldsymbol{\epsilon}|\mathbf{0}, {\sigma_0}^2\mathbf{I})$. Note that the columns of $\mathbf{W}$ correspond to the eigenvectors of the principal subspace, which we call eigen-bodyprints.

It can be demonstrated [13] that in PPCA the posterior distribution of the *latent features* can be expressed as:

$$p(\mathbf{z_n}|\mathbf{x_n}) = N(\mathbf{z_n}|\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x_n} - \boldsymbol{\mu}), {\sigma_0}^2\mathbf{M}^{-1}) \quad (2)$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma_0^2\mathbf{I}$. The previous equation is important because it is used to obtain maximum likelihood estimates of the *latent features* once the model parameters have been obtained.

The use of EM to obtain model parameters is straightforward. During the Expectation step, maximum likelihood estimates of the *latent features* are obtained using the current model parameters:

$$\begin{aligned} \mathbb{E}[\mathbf{z_n}] &= \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x_n} - \boldsymbol{\mu}) \\ \mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T] &= \sigma_0^2\mathbf{M}^{-1} + \mathbb{E}[\mathbf{z_n}]\mathbb{E}[\mathbf{z_n}]^T \end{aligned} \quad (3)$$

where $\boldsymbol{\mu}$ is the mean of the observed bodyprints. Then, model parameters are updated during the Maximization step:

$$\begin{aligned} \mathbf{W} &= \left[\sum_{n=1}^{N}(\mathbf{x_n} - \boldsymbol{\mu})\mathbb{E}[\mathbf{z_n}]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T]\right]^{-1} \\ \sigma_0^2 &= \frac{1}{ND}\sum_{n=1}^{N}\{\|\mathbf{x_n} - \boldsymbol{\mu}\|^2 - 2\mathbb{E}[\mathbf{z_n}]^T\mathbf{W}^T(\mathbf{x_n} - \boldsymbol{\mu}) \\ &\quad + Tr(\mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T]\mathbf{W}^T\mathbf{W})\} \end{aligned} \quad (4)$$

This process is repeated until convergence of the parameters. To evaluate the convergence, the EM algorithm uses the log-likelihood of the observed data as the objective function (see [44] for details).

PPCA with missing data. The problem of the EM algorithm as formulated in Equations 3 and 4 is that it cannot deal with missing values in the observed bodyprints (caused by occlusions). Fortunately, the EM provides a natural way to handle missing values. The three main differences that are found when dealing with missing data are:

- The mean of the observed bodyprints, $\boldsymbol{\mu}$, cannot be computed in a closed form and needs to be estimated in each iteration.
- The covariance matrix of the posterior distribution of the *latent features*, (Eq. 2), is different for each training sample since it depends on which variables of the bodyprint are observed in each sample.
- The formulation to obtain the transformation matrix $\mathbf{W}$ is more complex in this case, because each row of $\mathbf{W}$ needs to be calculated independently.

To indicate which bodyprint features have been observed, we use the binary matrix $\mathbf{O}$ so that $\mathbf{O}(m, n) = 1$ if the m feature of the training sample n is observed. Similarly, O_n is the set of indexes of the observed bodyprint features for sample n and O_m is the set of samples for which feature m has been observed. Using an element-wise formulation of Eqs 3 and 4, the EM steps can be modified to deal with missing data using only the columns of $\mathbf{W}$ and rows of $\mathbf{x_n}$ that correspond to the observed values (see [26] for a detailed explanation). In the Expectation step the latent features for each sample are estimated using the observed data and the current estimates of the parameters:

$$\mathbb{E}[\mathbf{z_n}] = \mathbf{M_n^{-1}}\sum_{m\in O_n}(\mathbf{x_n}(m) - \boldsymbol{\mu}(m))\mathbf{w_m} \quad (5)$$

$$\mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T] = \sum_{m\in O_n}(\sigma_0^2\mathbf{M_n^{-1}} + \mathbb{E}[\mathbf{z_n}]\mathbb{E}[\mathbf{z_n}]^T) \quad (6)$$

where $\mathbf{w_m}$ is the m^{th} column of $\mathbf{W}$ and $\mathbf{M_n} = \sum_{m\in O_n}\mathbf{w_m}\mathbf{w_m^T} + \sigma_0^2\mathbf{I}$. Note that $\mathbf{M_n}$ is different for each training sample. In the Maximization step, the model parameters are updated using the expected latent features:

$$\boldsymbol{\mu}(m) = \frac{1}{|O_m|}\sum_{n\in O_m}(\mathbf{x_n}(m) - \mathbf{w_m^T}\mathbb{E}[\mathbf{z_n}]) \quad (7)$$

$$\mathbf{w_m^T} = \left[\sum_{n\in O_m}(\mathbf{x_n}(m) - \boldsymbol{\mu}(m))\mathbb{E}[\mathbf{z_n}]^T\right]\cdot \left[\sum_{n\in O_m}\mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T]\right]^{-1} \quad (8)$$

$$\sigma_0^2 = \frac{1}{N}\sum_{m,n\in O}\{(\mathbf{x_n}(m) - \mathbf{w_m^T}\mathbb{E}[\mathbf{z_n}] - \boldsymbol{\mu}(m))^2 + \mathbf{w_m^T}{\mathbf{M_n}}^{-1}\mathbf{w_m}\} \quad (9)$$

Factor Analysis (FA). Factor Analysis is a latent variable model that is similar to PPCA. The main difference between them is that, in the generative model of Eq. 1, the noise distribution in FA is:

$$p(\boldsymbol{\epsilon}) = N(\boldsymbol{\epsilon}|0, \boldsymbol{\Psi}) \quad (10)$$

where $\boldsymbol{\Psi}$ is a general diagonal $l_x \times l_x$ matrix. The advantage of FA compared to PPCA is that it is a more flexible model that can capture different noise levels in the observed bodyprint features. The example of Figure 3 shows that this

situation can often appear in our re-identification context. Since $\mathbf{\Psi}$ is a diagonal matrix, the noise is also considered to be independent for each bodyprint feature in FA. This assumption is necessary to reduce the number of model parameters and avoid over-fitting.

The FA model can also be adapted to deal with missing bodyprint features using the EM algorithm. The equations for the algorithm are derived similarly as in PPCA, yielding the following expression for the Expectation step:

$$\mathbb{E}[\mathbf{z_n}] = \mathbf{G_n} \sum_{m \in O_n} \mathbf{w_m} \mathbf{\Psi}_m^{-1} (\mathbf{x_n}(\mathbf{m}) - \boldsymbol{\mu}(m)) \tag{11}$$

$$\mathbb{E}[\mathbf{z_n} \mathbf{z_n^T}] = \sum_{n \in O_m} (\mathbf{G_n} + \mathbb{E}[\mathbf{z_n}] \mathbb{E}[\mathbf{z_n}]^T) \tag{12}$$

where $\mathbf{G_n} = \sum_{m \in O_n} (\mathbf{I} + \mathbf{w_m} \mathbf{\Psi}_m^{-1} \mathbf{w_m^T})^{-1}$. For the Maximization step:

$$\mathbf{w_m} = \left[\sum_{n \in O_m} (\mathbf{x_n}(m) - \boldsymbol{\mu}(m)) \mathbb{E}[\mathbf{z_n}] \right] \left[\sum_{n \in O_m} \mathbb{E}[\mathbf{z_n} \mathbf{z_n^T}] \right]^{-1}$$

$$\mathbf{\Psi}_m = diag\{S_m - \\ - \mathbf{w_m} \frac{1}{|O_m|} \sum_{n \in O_m} \mathbb{E}[\mathbf{z_n}] (\mathbf{x_n}(m) - \boldsymbol{\mu}(m))^T\} \tag{13}$$

where $S_m = \sum_{n \in O_m} (\mathbf{x_n}(m) - \boldsymbol{\mu}(m))(\mathbf{x_n}(m) - \boldsymbol{\mu}(m))^T$.

Factor Analysis with known noise. Figure 4.b shows the graph representation of FA. It can be observed that $\mathbf{\Psi}$ remains constant for all samples and that it is inferred from the training samples as are the other model parameters. However, this can be a strong simplification because not all samples are equally corrupted by the noise. In this work, we propose a small modification to the FA model that is depicted in Figure 4.c. With this modification, it is possible to introduce different noise distributions for each sample. The noise distribution for each sample is a zero-mean multivariate Gaussian with diagonal covariance matrix $\mathbf{\Psi}_n$. Note that $\mathbf{\Psi}_n$ is placed in a blue circle, which means that the noise distribution is measured and introduced into the model. The main advantage of this new approach is that by introducing an estimate of $\mathbf{\Psi}_n$, we provide information about which features in each bodyprint are more reliable and which color features are less important (higher color variance over time). Note that since $\mathbf{\Psi}_n$ is different for each bodyprint, the information about the reliability of the features is different for each bodyprint.

If the noise parameters are provided to the algorithm, the only model parameters that need to be estimated during the training stage are the model mean $\boldsymbol{\mu}$ and the projection matrix $\mathbf{W}$. The EM equations used to obtain the model parameters in this case are the same as in the FA model except that $\mathbf{\Psi}_n$ no longer needs to be maximized (Eq. 13) and remains fixed (its value is provided as an input to the algorithm).

To obtain the noise distribution for each person, $\mathbf{\Psi}_n$, we use the corresponding temporal signatures (Figure 2). The idea is that, for each height, it is possible to extract the mean color over time together with the color variance since the color at each strip may vary over time. Again, to reduce the number of parameters, we assume that color features at different heights and color channels are independent ($\mathbf{\Psi}_n$ is diagonal).

4.3 Model discussion

The models introduced in Section 4.2 use the same probabilistic framework to extract the discriminant features. The probabilistic approach in combination with the EM iterative method to find the projection coefficients gives the algorithm the possibility to naturally handle features with missing data and to work with high dimensional feature vectors. The main difference among methods is based on how the noise covariance matrix is modeled. In PPCA, the noise is constant for all the samples and spatial dimensions. Therefore, this approach is indicated when there is no prior condition about noise. In contrast, the conventional FA method models the covariance matrix as a constant diagonal matrix, which means that each dimension may have different noise. In both cases, the covariance matrix is inferred during the EM iterations and directly affects the quality of the feature selection. On the other hand, the variant of FA with known noise considers that the noise is observed for each sample and dimension, so it does not need to be estimated throughout the EM loop. Therefore, it can be used to estimate the latent variables more accurately if the noise is correctly observed.

The experimentation in this paper will discuss which of the different approaches is more convenient for the purpose of person re-identification in multi-shot scenarios, where a person's appearance over time can be described as a unique, stable color feature vector with known variance.

5 Matching metric

The *latent feature* models introduced in Section 4 assume that the number of color features in the bodyprints are the same for all of the samples. More precisely, the height of each person is always quantized into 100 bins and a color feature of the bodyprint is obtained for each bin. Since the height of each person may be different, the quantization step is adapted accordingly. The idea behind using a fixed number of color features in the bodyprints is that we wanted to decouple appearance information from height information. In our preliminary experiments, we found that the height

normalization of bodyprints was very useful in reducing the dimensionality of the latent space because the variability of the training samples is reduced with the normalization. However, height information is also very important for our re-identification objective, and, for this reason, height information must be incorporated into the similarity metric.

Let $\mathbf{z}_i$ and $\mathbf{z}_j$ respectively be the *latent features* of person i and j obtained using any of the methods proposed in Section 4.2. The similarity $\mathcal{S}(i,j)$ between these two people is defined as follows:

$$\mathcal{S}(i,j) = \mathcal{M}(\mathbf{z}_i, \mathbf{z}_j) e^{-\frac{\Delta H}{\nu}} \quad (14)$$

where the first term measures the appearance similarity using the *latent features* and the second term penalizes the difference in height, ΔH, between the two people. The constant ν controls how the confidence decreases and its value was empirically determined (ν = 4cm). In our experiments, several appearance measurements for $\mathcal{M}(\mathbf{z}_i, \mathbf{z}_j)$ are compared. Specifically, we evaluated Euclidean, cosinus, Euclidean-Mahalanobis, and cosinus-Mahalanobis distances. Details about the implementations of these measurements can be found in [14].

6 Evaluation

In this section, we evaluate the proposed *latent feature* models in the context of person re-identification. Information about the dataset on which the experiments are carried out is provided in Section 6.1. Section 6.2 describes the system evaluation methodology. Finally, the re-identification results are presented in Section 6.3.

6.1 Dataset description

The re-identification results presented in this work were obtained using data recorded in a real store. The people in the videos were not aware of the cameras, and therefore they behave naturally. The raw recorded data is publicly available at [2] so that other researchers can evaluate their algorithms with a common dataset. Although we used all of the available frames for each person to extract the features, other researchers might use different schemes, such as key frame selection algorithms or even neglect depth information. For this reason, we thought that it was important to deliver the raw data together with the masking information for each person so that different schemes can be tested and compared.

In our recording set-up, two different and non-overlapping RGB+depth cameras covering entrance and exit areas were installed. The cameras were hanging on a wall 3 meters above the ground and pointing 30 degrees downwards. Since cameras were placed at different locations, illumination changes were likely to occur. A few examples of the people in the dataset can be seen in Fig. 8. In each row, the same person can be seen at the entrance and at the exit. The dramatic changes in pose and view angles can be observed. It can also be observed that the people are free to carry objects, shopping baskets, or push shopping carts. In total, the dataset contains 73 different people, where each person is tracked at least during 7 frames and a mean of 15 frames. The image size is 320x240p and the frame rate of image acquisition is 7 fps for each camera.

6.2 Evaluation methodology

To evaluate the performance of our system, we used the re-identification rate and the average cumulative match characteristic (CMC) curves [24] over 40 trials. The re-identification rate shows the percentage of correct matches. The CMC curve represents the re-identification probability given that the good match is in the first r ranked candidates. The re-identification rate is a particular case of CMC with $r = 1$.

In the experiments, we analyzed the contribution of the probabilistic latent variable models presented in Section 4.2 applied to color features such as bodyprints, and we compared the results against state-of-the-art works such as SDALF [18] and PRDC [49]. Although these methods do not use depth information and are therefore at a disadvantage in the comparison, we just wanted to compare them to show the real contribution of this work as a robust ensemble for identification. Note that none of the works used in the comparison use any other contextual cues such as temporal causality (a person at the exit must first have entered). Even though this type of contextual information is very valuable because it reduces the effective search population in a real scenario, we preferred to ignore it so that the results reflect only the re-identification abilities of the algorithms.

The proposed models (Fig.4) were trained using the bodyprints of the people at the entrance. Once the models were trained, the *latent features* of all bodyprints were extracted. Finally, the matching was conducted using the metrics described in Section 5.

6.3 Results

Similar to PCA, the columns of the projection matrix W can be interpreted as the directions in the observed feature space that capture more variance of the training data. The maximum number of *latent features* that can be obtained is limited by the number of training samples in a dataset (in our case 73). Usually, latent features that capture more variance are considered to be more relevant since they contribute more to the reconstruction of the observed feature space. In most pattern recognition problems, the number of *latent features* to be considered is a trade-off between capturing as

much variance as possible and discarding low variance features that capture the noise of the training set.

In this work, we studied the influence of the number of *latent features* for all the models and metrics proposed in the paper. As mentioned above, the evaluation was carried out using the re-identification rates for each case. Figures 5, 6, and 7 show the re-identification rates for all of the proposed cases. The first general conclusion is that a few *latent features* are enough to extract the relevant information in all of the cases, which confirms our initial hypothesis about using dimensionality reduction techniques. Another result is that the Euclidean metric is the one that attains the best performance for the three latent models.

The PPCA model achieved a re-identification rate of 62.5% using a small number of *latent features* (a number between 20 and 30 components). In the case of FA with unknown variance, the best performance was a little bit smaller (61.1%.) although the performance never decreased with more *latent features*. In the case of FA with known noise, the best re-identification rate was also 62.5% using 40 *latent features*. Even though the best rate in FA with known noise equaled the best rate of PPCA, the results in general were worse than in PPCA. This general lower performance in FA is due to the fact that the noise may not be accurately estimated, since we only considered the variation of the mean color per strip over time, but we did not take into account the noise introduced in the estimation of the mean color at each strip at a single frame.

Figure 8 shows several correct matching examples using 20 PPCA *latent features*. The left and right columns show the images obtained at the entrance and exit, respectively. The persons that are matched in each case have been surrounded by an ellipse. Note that in our dataset we do not impose any restriction on the person's behaviour, so in the second example the old woman is pushing a shopping cart at the entrance but the cart does not appear at the exit. These examples show how our algorithm can effectively deal with big changes in pose and also with outliers produced by carried objects. For instance a shopping basket in the first example and a shop item in the third example. In contrast, Fig. 9 shows a few examples where re-identification failed. The examples clearly show the difficulty of our dataset where in many cases people are wearing similar clothes and re-identification is even difficult for the human eye.

In Table 2, we compare the re-identification rates achieved using the *latent features* with our previous work using explicit features or bodyprints [3]. It can be observed that the use of *latent features* has significantly improved the performance, which is mainly due to the fact that the *latent features* have naturally chosen the discriminative features that contain the relevant information.

Figure 10 shows a comparison of the performance of the *latent features* that is calculated using PPCA with 20 components versus SDALF [18] and PRDC [49]. The figure also shows the benefit of using the height difference factor of Eq. 14. As can easily be seen, our method overtakes the reference methods especially for the lower ranks in the CMC curve. This difference in the results is caused by two mentioned factors: 1) the use of depth information allows a person appearance to be described taking into account their real height under a frontal perspective obtained by applying the inverse camera projection; 2) the use of latent features allows missing data to be handled, removes noise, finds the discriminant features, and provides a useful tool for working with the temporal signature of a person over time by compressing all the different appearances into a stable feature vector plus a variance that is absorbed by the latent model during training.

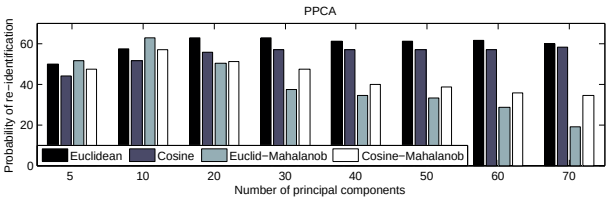

Fig. 5 Probability of detection (rank $r = 1$) using PPCA for different numbers of principal components.

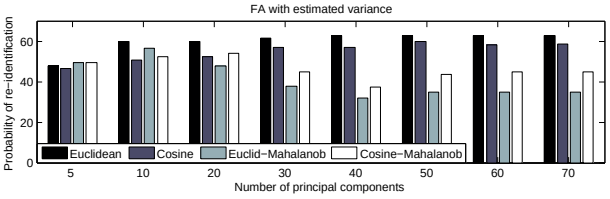

Fig. 6 Probability of detection (rank $r = 1$) using FA with unknown variance for different numbers of principal components.

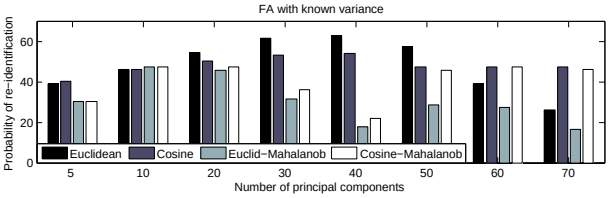

Fig. 7 Probability of detection (rank $r = 1$) using FA with known variance for different numbers of principal components.

Table 2 Summary of the best re-identification rates for the proposed methods

Feature type	Re-identification Rates
Explicit Appearance Features [3]	52.5%
PPCA	62.5%@20 components
FA unknown variance	61.5%@40 components
FA known variance	62.5%@40 components

Fig. 8 Examples of correct matches using the PPCA method with 20 components (rank $r = 1$). The left and right columns show people at the entrance and exit of the shop, respectively.

Fig. 9 Examples of incorrect matches using the PPCA method with 20 components (rank $r = 1$).

7 Conclusions

In this paper, we have introduced the concept of *latent features* for person re-identification. *Latent features* are extracted from an explicit appearance descriptor (named bodyprint) using different probabilistic latent variable models. The use of *latent features* minimizes some problems such as outliers, noise, missing data, and correlation of bodyprint features.

In this work, different latent variable models have been proposed and compared. The basic difference among them is how noise is handled in each case. The results show that the best re-identification rates are obtained using PPCA, although FA with unknown variance provides more stable results since the re-identification rate does not decrease when more *latent features* are used. Compared to our previous work, which is based on bodyprints and the state-of-the-art methods for person re-identification, the use of *latent features* significantly increases the global performance.

One of the current problems of our system is that bodyprints assume a Gaussian distribution for the color features at each height. However, this is too strong an assumption when outliers are present or when the color distribution is multimodal (clothes with colored stripes). For this reason, our future work will focus on the use of new appearance models that can cope with this multi-modality and on how *latent features* can be used in this context.

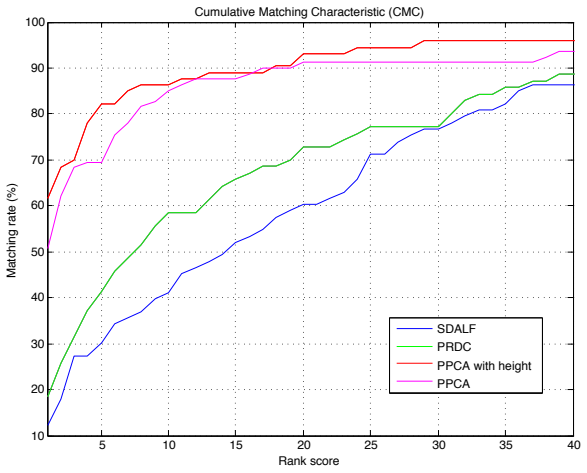

Fig. 10 Performance comparison using CMC curves

Acknowledgements The work presented in this paper has been funded by the Spanish Ministry of Science and Technology under the CICYT contract TEVISMART, TEC2009-09146.

References

1. http://kinectforwindows.org/
2. http://www.gpiv.upv.es/videoresearch/personindexing.html
3. Albiol A., Albiol Al., Oliver J., Mossi J.M.: Who is who at different cameras. matching people using depth cameras. IET Comput. Vis. (2012)
4. Bak S., Corvee E., Bremond F., Thonnat M.: Person re-identification using haar-based and dcd-based signature. In: 2nd

Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, AMMCSS 2010, in conjunction with 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. AVSS (2010)
5. Bak S., Corvee E., Bremond F., Thonnat M.: Person re-identification using spatial covariance regions of human body parts. In: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 435–440 (2010)
6. Bak S., Corvee E., Bremond F., Thonnat M.: Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In: Advanced Video and Signal-Based Surveillance. Klagenfurt, Autriche (2011). URL http://hal.inria.fr/inria-00620496
7. Baltieri D., Vezzani R., Cucchiara R., Utasi A., BenedeK C., Szirányi T.: Multi-view people surveillance using 3d information. In: ICCV Workshops, pp. 1817–1824 (2011)
8. Barbosa, B.I., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: First International Workshop on Re-Identification (2012)
9. Basilevsky A.: Statistical factor analysis and related methods: theory and applications. Willey (1994)
10. Bäuml M., Bernardin K., Fischer k., Ekenel H.K., Stiefelhagen R.: Multi-pose face recognition for person retrieval in camera networks. In: International Conference on Advanced Video and Signal-Based Surveillance (2010)
11. Bazzani L., Cristani M., Perina A., Farenzena M., Murino V.: Multiple-shot person re-identification by hpe signature. In: Proceedings of the 2010 20th International Conference on Pattern Recognition, pp. 1413–1416. Washington, DC, USA (2010)
12. Bird N.D., Masoud O., Papanikolopoulos N.P., Isaacs A.: Detection of loitering individuals in public transportation areas. Intelligent Transportation Systems, IEEE Transactions on **6**(2), 167–177 (2005)
13. Bishop C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
14. Cha S.H.: Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences **1**(4), 300–307 (2007)
15. Cheng Y.M., Zhou W.T., Wang Y., Zhao C.H., Zhang S.W.: Multi-camera-based object handoff using decision-level fusion. In: Conference on Image and Signal Processing, pp. 1–5 (2009)
16. Dikmen M., Akbas E., Huang T.S., Ahuja N.: Pedestrian recognition with a learned metric. In: Asian Conference in Computer Vision (2010)
17. Doretto G., Sebastian T., Tu P., Rittscher J.: Appearance-based person reidentification in camera networks: Problem overview and current approaches. Journal of Ambient Intelligence and Humanized Computing pp. 1–25 (2011)
18. Farenzena M., Bazzani L., Perina a., Murino V., Cristani M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010). IEEE Computer Society, San Francisco, CA, USA (2010)
19. Fodor I.: A survey of dimension reduction techniques. Tech. rep., Lawrence Livermore National Laboratory (2002)
20. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. In: J. Mach. Learn. Res., vol. 4, pp. 933–969 (2003)
21. Gandhi T., Trivedi M.: Panoramic appearance map (pam) for multi-camera based person re-identification. Advanced Video and Signal Based Surveillance, IEEE Conference on **0**, 78 (2006)
22. Garcia, J., Gardel, A., Bravo, I., Lazaro, J.: Multiple view oriented matching algorithm for people reidentification. Industrial Informatics, IEEE Transactions on **10**(3), 1841–1851 (2014)
23. Gheissari N., Sebastian T.B., Hartley R.: Person reidentification using spatiotemporal appearance. In: CVPR (2), pp. 1528–1535 (2006)
24. Gray D., Brennan S., Tao H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS) (2007)
25. Gray D., Tao H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the 10th European Conference on Computer Vision: Part I, pp. 262–275. Berlin, Heidelberg (2008)
26. Ilin A., Raiko T.: Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. **99**, 1957–2000 (2010)
27. Javed O., Shafique O., Rasheed Z., Shah M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Comput. Vis. Image Underst. **109**(2), 146–162 (2008)
28. Kai J., Bodensteiner, C., Arens, M.: Person re-identification in multi-camera networks. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, pp. 55 –61 (2011)
29. Kuo C.H., Huang C, Nevatia R.: Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In: Proceedings of the 11th European conference on Computer vision: Part I, ECCV'10, pp. 383–396. Springer-Verlag, Berlin, Heidelberg (2010)
30. Lan, R., Zhou, Y., Tang, Y.Y., Chen, C.: Person reidentification using quaternionic local binary pattern. In: Multimedia and Expo (ICME), 2014 IEEE International Conference on, pp. 1–6 (2014)
31. Loy, C.C., Liu, C., Gong, S.: Person re-identification by manifold ranking. In: icip, pp. 3318–3325 (2013)
32. Madden C., Cheng E., Piccardi M.: Tracking people across disjoint camera views by an illumination-tolerant appearance representation. Machine Vision and Applications **18**, 233–247 (2007)
33. Mazzon R., Tahir S.F., Cavallaro a.: Person re-identification in crowd. Pattern Recognition Letters **33**(14), 1828–1837 (2012)
34. Oliveira I.O., Souza Pio J.L.: People reidentification in a camera network. In: Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 461–466 (2009)
35. Papadakis P., Pratikakis I., Theoharis T., Perantonis S.J.: Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. International Journal of Computer Vision **89**(2-3), 177–192 (2010)
36. Prosser B., Zheng W.S., Gong S., Xiang T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference, pp. 21.1–21.11. BMVA Press (2010)
37. Roweis S.: Em algorithms for pca and spca. In: in Advances in Neural Information Processing Systems, pp. 626–632. MIT Press (1998)
38. S., P., J., O., S., V., B., B.: Local fisher discriminant analysis for pedestrian re-identification. In: CVPR, pp. 3318–3325 (2013)
39. Satta R., Fumera G., Roli F.: Fast person re-identification based on dissimilarity representations. Pattern Recognition Letters, Special Issue on Novel Pattern Recognition-Based Methods for Reidentification in Biometric Context **33**, 1838–1848 (2012)
40. Tao, D., Jin, L., Wang, Y., Li, X.: Person reidentification by minimum classification error-based kiss metric learning. Cybernetics, IEEE Transactions on **45**(2), 242–252 (2015)
41. Tipping M.E., Bishop C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B **61**, 611–622 (1999)
42. Tisse C.L., Martin Lionel, Torres Lionel, Robert Michel: Person identification technique using human iris recognition. In: Proc. of Vision Interface, pp. 294–299 (2002)
43. Vandergheynst P., Bierlaire M., Kunt M., Alahi A.: Cascade of descriptors to detect and track objects across any network of cameras. Computer Vision and Image Understanding pp. 1413–1416 (2009)

44. Verbeek, J.: Notes on probabilistic pca with missing values. Tech. rep., Tech. report (2009)
45. Wang D., Chen C.O., Chen T.Y., Lee C.T.: People recognition for entering and leaving a video surveillance area. In: Fourth International Conference on Innovative Computing, Information and Control, pp. 334–337 (2009)
46. Zhang Z, Troje N.F.: View-independent person identification from human gait. Neurocomputing **69**, 250–256 (2005)
47. Zhao T., Aggarwal M., Kumar R., Sawhney H.: Real-time wide area multi-camera stereo tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 976–983 (2005)
48. Zheng, S., Xie, B., Huang, K., Tao, D.: Multi-view pedestrian recognition using shared dictionary learning with group sparsity. In: B.L. Lu, L. Zhang, J.T. Kwok (eds.) ICONIP (3), *Lecture Notes in Computer Science*, vol. 7064, pp. 629–638. Springer (2011)
49. Zheng W.S., Gong S., Xiang T.: Person re-identification by probabilistic relative distance comparison. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 649–656 (2011)