Who is who at different cameras. People re-identification using Depth Cameras

Antonio Albiol , Alberto Albiol, Javier Oliver, José Manuel Mossi

Universitat Politecnica Valencia

Camino Vera s/n, 46022 Valencia (Spain)

aalbiol@dcom.upv.es , alalbiol@iteam.upv.es

jaolmol@upvnet.upv.es , jmmossi@dcom.upv.es

July 17, 2012

Abstract

This paper proposes the concept of bodyprints to perform re-identification of people in surveillance videos. Bodyprints are obtained using calibrated depth-color cameras such as kinect. Our results on a database of 40 people show that bodyprints are very robust to changes of pose, point of view and illumination. Potential applications include tracking people with networks of non-overlapping cameras.

1 Introduction

The use of camera networks has widespread in many different domains such as surveillance in large facilities (airports, bus/train stations), sports/conference venues and retail stores, just to enumerate a few. According to a recent survey [17], only a small fraction of these videos is ever watched or analyzed. This fact raises the need of developing high level video analysis tools to obtain relevant information.

For example, in the context of retail, information about the statistical analysis of people behaviors is very useful for marketing purposes. Aspects such as frequent paths, locations where people tend to stop,

people flow at certain locations etc. are examples of the kind of information that is interesting for a shop manager in order to measure the impact of his decisions.

An important problem that appears in multi-camera systems is to re-identify people that leave one camera and enter in another or in the same camera after a period of time. Re-identification is a key aspect for tracking applications in non-overlapping cameras scenarios. One possible application of re-identification, and the one that motivated our research, is to determine how much time do people spend in a shop by analyzing images of their entry and departure taken by one or more cameras.

The recent appearance of low cost RGB-Depth (RGB-d) cameras, such as Microsoft Kinect offers the opportunity to explore the use of many three-dimensional computer vision techniques in surveillance scenarios at an affordable cost. In RGB-d cameras, each pixel has an associated depth, apart from its RGB color value. This information allows to segment objects more easily, even in presence of severe occlusions, using only the spatial coordinates of each pixel. In fact, a new trend of processing RGB-d images using point clouds rather than regularly arranged pixels has just emerged [1].

In this paper, we will concentrate in the re-identification problem using Kinect cameras. The key idea is to take advantage of the available spatial information to obtain a feature vector per person, which we call bodyprint. Bodyprints can be matched across different cameras or the same camera at a later time to solve the re-identification problem.

The rest of the paper is organized as follows. Section 2 reviews the state of the art on video re-identification. In section 3.1 we will briefly describe the sensor. In section 3.2 the scenario calibration procedure that has been used is presented. People segmentation and tracking are presented in sections 3.3 - 3.5. Bodyprints, the characteristic signatures that will allow to match people are described in section 4.1. Section 4.2 will address the metric used to compare bodyprints. Finally we will provide results and some conclusions about what can be expected from the technique presented in this paper.

2 Related research

The basic assumption of most people re-identification approaches is that people are wearing the same clothes in all the views. With this assumption, the problem is how to model the global appearance of individuals when large variations in pose, point of view and illumination are expected. The use of traditional biometric modalities such as face is generally limited because it is quite common that cameras are set up to cover a relatively large area and generally there is not enough resolution to perform face recognition. One of the few approaches that uses faces is presented in [14]. In this work the cameras are set up to cover a narrow area in corridors. This particular configuration allows to increase the resolution of faces and extract face tracks which can be matched against a database. Another example that uses faces is presented in [15]. In this work the face modality, if available, is combined with other global appearance modalities, however the face modality is only available in 5% of the matches.

Gait has been also proposed by several authors as a promising cue for people re-identification [29, 15, 13]. In this case, the difficulty of extracting reliable gait features in unconstrained multicamera scenarios may explain why this modality has not been widely used for people re-identification.

Appearance methods for people re-identification are commonly grouped into single or multiple-shot [9]. Single-shot methods use only one image to perform identification, while multiple-shot methods use different images of the same person obtained by tracking. Images from each individual are usually taken using one camera, although in [20] a multiple camera configuration that allows to simultaneously capture individuals from several points of view is proposed.

The models used to describe people appearance can be also classified into holistic and part-based. The key difference between these methods is that part-based algorithms extract different features for each body part, such as legs, trunk and head, and feature comparison is done accordingly to this body segmentation. Although part-based methods are very promising, holistic methods are still more robust in challenging scenarios [19]. A couple of representative part-based methods are [10, 28].

One common approach to describe color appearance is to use color histograms. For instance, in [9] the color global histogram is combined with recurrent local patterns, that characterize texture, after

epitomic analysis. Raw color features are not very robust because they highly depend on the illumination conditions and different camera color sensitivity. For this reason, it is very important to use some sort of color normalization. An example of color normalization is found in [8]. In this work, colors are quantized to create a palette of just eleven basic colors.

The problem associated with color histograms is that spatial information is discarded. To avoid this problem, in [18] the image is divided into a grid of fixed cells and then, color histograms are obtained for each cell. A different option is to divide the image of the pedestrian into horizontal stripes [12] and then obtain color features for each stripe.

In addition to color information, there are many systems that also use texture information to identify people. It is quite common that texture is obtained around interest points. For instance, in [16] SURF points are used to match people and in [24], interest points are extracted using a combination of SIFT and SURF algorithms. A different approach to extract local features is presented in [22] where machine learning is used to extract the most discriminative feature set.

Fair comparison of different person identification approaches requires using a common data. Two of the most widely datasets used in the literature include the ViPER and i-Lids datasets. The ViPER [21] database is specifically designed for single-shot identification, and contains images of many individuals under large variations of illumination and pose. The i-Lids database [25] was recorded in an airport arrival halls at busy hours. In this real scenario, occlusions are a common problem that was not present in the ViPER dataset. Unfortunately, none of the public datasets contains depth information and for this reason we have created a new dataset which contains RGB-d images and made it publicly available [2].

3 System description

Figure 1 shows the block diagram of the proposed system. Using kinect depth information and scene calibration, the height map block creates a virtual cenital view of the scene. This particular point of view greatly simplifies the segmentation and allows to deal with situations of severe occlusions. Segmented objects are tracked in the scene to obtain as much information as possible. This is important because due

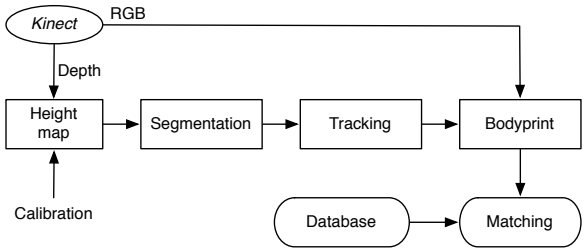

Figure 1: Block diagram of the overall system.

to occlusions, not all body parts are observed in all the frames. Finally, bodyprints are obtained for each person using RGB information. Bodyprints can be matched against previously recorded bodyprints in the same or a different location. Next subsections will describe each of these blocks in detail.

3.1 The kinect sensor

The kinect sensor was introduced by Microsoft as an input device for its game console Xbox-360. When used for gaming, it segments people and is able to estimate their pose and actions in order to command games. Shortly after its commercialization, PrimeSense [3] (the company that developed Kinect) released a library called OpenNI [4] to ease the development of interective applications. OpenNI provides an easy API to get access to RGB and depth information and now, it is also used by other well known open source projects such as OpenCV (starting at version 2.2) [5], ROS [6] and PCL [1]. Recently, Microsoft Research released an SDK with identical functionality as OpenNI [7].

The kinect sensor has a color RGB and an infrared (IR) cameras. It also has an IR pattern generator, that jointly with the infrared camera is able to determine the depth. A limitation of this sensor is that it can only work in indoors environments under normal illumination conditions. In outdoors, the IR pattern has not enough contrast to measure distance.

The device can be configured to *align* both images (RGB and depth). This is required because both cameras have a different focal length and are located at slightly different positions. An example of RGB image and its associated depth is shown in Fig. 2. For a number of reasons, depth information may not be available for all pixels. In that figure, locations where no depth information is available are shown in black. RGB and depth images are aligned taking the RGB as reference. The sensor has a minimum

Figure 2: Examples of aligned RGB and depth images obtained with kinect.

distance to measure depth of around one meter, and a maximum distance of about 10 m. These values do not represent any problem for many indoor surveillance scenarios.

3.2 Camera calibration

Using depth information and intrinsic kinect parameters, the spatial camera coordinates of the i-th pixel can be determined using the following equations:

$$\begin{aligned} z^i_{cam} &= d_i \\ x^i_{cam} &= (x_i - \bar{x})\, z^i_{cam}/f \\ y^i_{cam} &= (y_i - \bar{y})\, z^i_{cam}/f \end{aligned}$$

where f is the focal length (in pixels), x_i and y_i are the pixel coordinates in the image (in pixels), d_i is the depth associated with the pixel, and $\bar{x}$ and $\bar{y}$ are the coordinates of the image center. Previous equations provide 3D coordinates in a coordinate system that has the origin at the optical center of the camera, and that it is aligned with the camera axis (see figure 3).

The goal of camera calibration is to extract the transformation parameters that allow to change from camera coordinates to world coordinates. We are seeking a coordinate system where the $\hat{Z}_{world}$ axis is

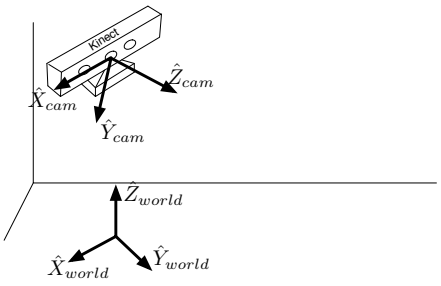

Figure 3: Relation between camera and world coordinates.

aligned with the true vertical axis and with its origin located on the ground plane. This way, the z_{world} coordinate represents the height with respect to the ground plane. Figure 3 shows the camera and world coordinate systems.

The rigid transformation that maps both coordinate systems can be written as:

$$\begin{pmatrix} x_{world} \\ y_{world} \\ z_{world} \end{pmatrix} = \mathbf{R} \begin{pmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \end{pmatrix} + \mathbf{T} \tag{1}$$

Some full automatic methods exist to determine the ground plane [27]. However, we have used a simple approach that only requires to manually select a portion of the RGB-d image that corresponds to the ground plane (see Fig. 4).

The method starts by computing the camera coordinates of every point under the mask (for which depth data is available). The ground points are assumed to be on a plane in space, therefore the eigenvector associated to the smallest eigenvalue of the covariance matrix of the ground points will give the direction corresponding to the normal of the plane. Let's call this direction $\hat{Z}_{world}$. At this point, we only know that $\hat{Z}_{world}$ is normal to the ground plane, but we do not know yet if it is pointing upwards or downwards. To resolve this ambiguity, we compute the projection of each ground point to the $\hat{Z}_{world}$ direction:

$$z^i_{world} = \vec{P}^i_{cam} \cdot \hat{Z}_{world} \tag{2}$$

where $\vec{P}^i_{cam} = (x^i_{cam}, y^i_{cam}, z^i_{cam})^t$ and '$\cdot$' denotes the dot product. We are interested in getting a unit vector $\hat{Z}_{world}$ that points *upwards*. The mean value of z^i_{world} of the pixels under the ground mask is

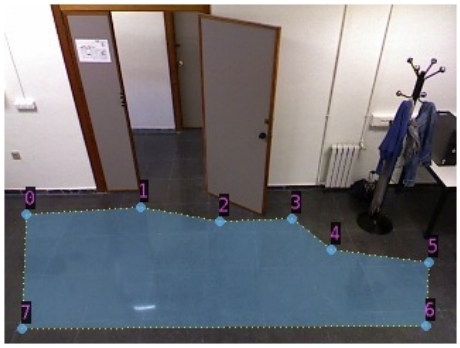

Figure 4: Portion of ground used for calibration.

the relative height of ground with respect to camera. If this value is negative, it means that ground is lower than camera. This is the normal situation as shown in Figure 3, and in this case $\hat{Z}_{world}$ is actually pointing upwards. In the opposite case, we just change its direction, $\hat{Z}_{world} \leftarrow -\hat{Z}_{world}$.

The choice of the $\hat{X}_{world}$ and $\hat{Y}_{world}$ axis is somewhat arbitrary (it does not affect the segmentation), provided that they are orthogonal to $\hat{Z}_{world}$ and between them. Without loss of generality, we have chosen $\hat{X}_{world}$ to be approximately aligned to $\hat{X}_{cam}$ and orthogonal to $\hat{Z}_{world}$:

$$\hat{X}_{world} = (1,0,0)^T - ((1,0,0)^T \cdot \hat{Z}_{world})\, \hat{Z}_{world}$$

and then normalizing for unit norm:

$$\hat{X}_{world} \leftarrow \hat{X}_{world}/|\hat{X}_{world}|$$

For $\hat{Y}_{world}$ we simply compute the cross product between $\hat{Z}_{world}$ and $\hat{X}_{world}$

$$\hat{Y}_{world} = \hat{Z}_{world} \times \hat{X}_{world}$$

The unit column vectors $\hat{X}_{world}$, $\hat{Y}_{world}$ and $\hat{Z}_{world}$ determine the rotation matrix as:

$$\mathbf{R} = \begin{pmatrix} \hat{X}_{world}^T \\ \hat{Y}_{world}^T \\ \hat{Z}_{world}^T \end{pmatrix}$$

Finally, we need to determine the components of the translation vector $\mathbf{T}$. The t_x and t_y components are irrelevant for the segmentation process, and we have set them to be zero (the world coordinate origin lies

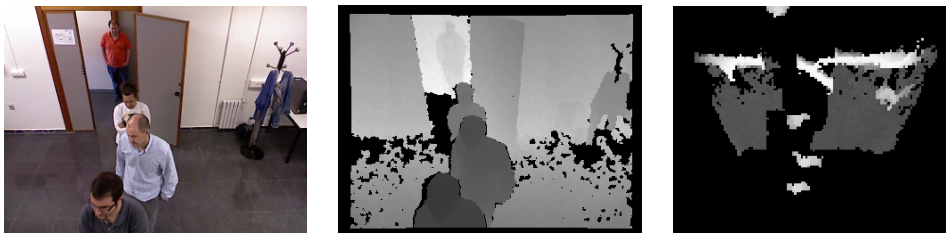

Figure 5: Height Map example and corresponding image and depth.

just under the camera as shown in Fig 3). To value of t_z has been chosen so that ground points have a height of around zero:

$$t_z = -\text{mean}\left\{z^i_{world}\right\}$$

$$\mathbf{T} = \begin{pmatrix} 0 \\ 0 \\ t_z \end{pmatrix}$$

3.3 Height Maps

The previous calibration allows to create a virtual cenital view of the scene which we call height maps. Height maps are images where the pixel values represent the height with respect to the ground. The use of height maps makes the segmentation of the scene much simpler. This idea has also been proposed in other works [30].

The process used to obtain height maps starts by representing all pixels for which depth information is available with world coordinates using the transformation of eq. 1. Then, the plane XY_{world} is quantized in small bins of size 5×5 cm. Pixel i is assigned to a bin according to its x^i_{world} and y^i_{world} coordinates. Finally, each bin is represented with the maximum height (max. z^i_{world}) of its assigned points. Figure 5 shows an example of height map. The origin (position of the camera) is at the bottom center of the height map. Upwards in the map means getting further from the camera ($\hat{Y}_{world}$ axis). Horizontal axis on the map corresponds to $\hat{X}_{world}$. On the map, black means a bin for which there has been no point falling into it (mostly due to occlusions). Dark grey means a height around zero (ground plane) while brighter intensity means higher with respect to the ground.

3.4 Segmentation

From the example in Fig. 5, it can be seen that people appear as clearly separated bright blobs in the height map. Thresholding the height map followed by connected component analysis would yield an almost perfect discrimination of people except in the case that two people are in contact. To cope with this situation the following segmentation algorithm has been used:

1. Mask out the location of walls and other fixed furniture.

2. Threshold the height map at an appropriate level. We have used 120 cm. above the ground plane as a threshold.

3. Group the pixels in thresholded height map into connected components.

4. Each connected component (blob) may correspond to one or more people. If the blob is large enough we try to split it. To do so we observe that heads are local maxima of the height map. To prevent over-segmentation, the idea of contrast of a local maximum is borrowed from mathematical morphology [23, 26]. We remind that the contrast of a regional maximum is *the minimum descent in order to reach a higher maximum*. Figure 6 shows a detail of a blob corresponding to two people where two contrasted maxima are clearly visible. For each connected component determine the contrast of the regional maxima and select those that have a contrast of at least 20 cm.

5. If more than one regional maximum with enough contrast is found within a blob, we will subdivide it using the watershed algorithm using the valid maxima as markers [11].

6. The number of image pixels associated with each local maxima is computed and those that do not contain enough (50 in all our tests) image pixels are discarded. This is to avoid spurious small objects created by distance noise.

7. Finally, pixels are labelled according to the label of their corresponding bin in height map.

An example of segmentation result is depicted in Fig. 7. A few remarks are pertinent at this point:

- Segmentation provides us information of which pixels in the original image belong to each particular person.

- Remember that for each pixel in the original image we know its height (z^i_{world}) and its RGB value.

- An estimate of the height of each person is readily available from the segmentation as the maximum value with a given label in the height map. This can be useful as a first clue in the process of matching people at entry and exit (only people of similar height will be tested).

- The position of the i-th person on the height map, p_i, will be the center of gravity of the blob.

3.5 Tracking

Tracking is the process of linking the segmentation results from several frames. Henceforth, a *track* will denote a thread of linked objects corresponding to the same person. Since quality of the segmentation is very good and the frame rate is relatively high compared to the motion of people, a quite simple tracking algorithm has been used.

Let's assume that N_t people are detected at t, and that in $t-1$ we had N_{t-1} tracks. For each track in $t-1$, we predict its position in t assuming that the displacement on the XY plane between $t-1$ and t will be the same as between $t-2$ and $t-1$.

Then, we compute a $N_{t-1} \times N_t$ distance matrix in which the elements contains the euclidean distances between the people positions of instant t and the N_{t-1} predicted track positions. Using the distance matrix, we track objects by repeatedly:

- Find the position of the minimum of the distance matrix.

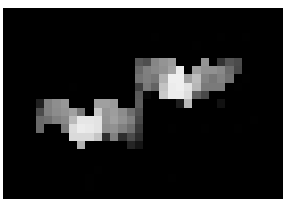

Figure 6: Portion of height map showing the presence of two contrasted maxima due to two heads.

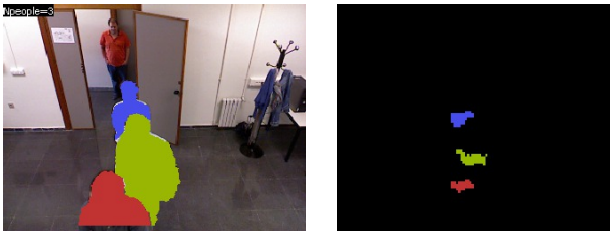

Figure 7: Segmentation results of example in Fig. 5.

- If this distance is below a suitable threshold, the corresponding object and track are linked and the corresponding row and column removed. Notice that this maximum distance (in metres) is related to frame rate and the typical walking speed of a person.

- When no more objects can be linked the loop ends and we start a new track for each un-matched object, and terminate all the un-matched tracks.

A few tracking examples can be viewed in [2]. In the video, it is possible to see the label assigned to each person and the height map used for segmentation.

4 Bodyprints

In order to match people, we extract a feature vector per track which we call *bodyprint*. Bodyprints act in a similar way to fingerprints. They are a sufficiently distinctive set of features that allow to discriminate people.

4.1 Extraction

The key idea of bodyprints is that each of its elements summarizes the color appearance at a different height for a track. Our algorithm is similar to [12] that also divides a person into stripes at different heights. However, our approach takes advantage of the good precision of the kinect sensor and the scene calibration that allows to precisely compute the height of all the pixels associated to a person track. The use of all the pixels associated to a track is important because, due to occlusions, not all body parts are visible in all frames.

Height h is discretized at steps of 2cm. At time t, we compute the mean RGB value, for each given

height, of the pixels that belong to the k-th track and are at $Z_{world} = h$:

$$\overline{RGB}_k(t, h)$$

Fig 8 shows a few examples of temporal signatures $\overline{RGB}_k(t, h)$ where the horizontal axis represents time and the vertical height. Black values indicate that no pixels were found at that particular height and instant. Note that in these examples people are walking downwards and for this reason the feet are the first part that leaves the image (right portion of temporal signature). Notice also that the width of the signatures varies because it depends on the time that a particular person was visible (which in turn may depend on the person walking speed). The temporal signatures exhibit a small ripple caused by variations of height while walking. This ripple can be easily compensated, however our experiments show that the impact of this artifact in the final performance is almost negligible. The count of pixels that contribute to each value of $\overline{RGB}_k(t, h)$ will be denoted as $\mathcal{C}_k(t, h)$.

Finally, we obtain bodyprints by averaging the temporal signatures along time:

$$\overline{RGB}_k(h) = \frac{1}{\mathcal{C}_k(h)} \sum_t \overline{RGB}_k(t, h)\, \mathcal{C}_k(t, h) \tag{3}$$

$$\mathcal{C}_k(h) = \sum_t \mathcal{C}_k(t, h) \tag{4}$$

It can be seen that all the information from a person has been compacted into a bodyprint vector, $\overline{RGB}_k(h)$, and a count vector, $\mathcal{C}_k(h)$. The bodyprint vector describes the appearance of the person. The count vector is a measure of the reliability of the values in the bodyprint.

Just to illustrate what happens in the case of a temporary occlusion, in Fig. 9 we can see a temporal signature containing a brief occlusion. Notice the missing information during the occlusion. However, thanks to the tracking, it is possible to build the whole bodyprint merging the information of several instants. In Fig. 9 another kind of occlusion is also shown. The man with the blue shirt is partially occluded all the time but since we have information for every height, the bodyprint is fully formed. On the same scene, the lady closer to the camera is only visible in her upper part. This would yield a bodyprint with some missing portions (the lower part). Matching bodyprints containing missing portions is possible as it will be explained in next section.

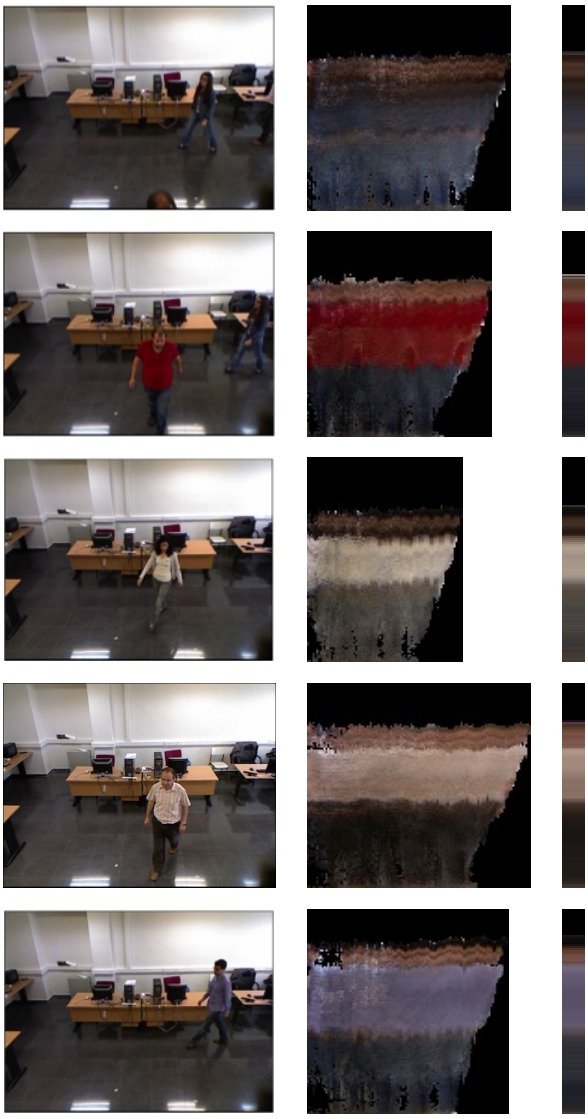

Figure 8: Left: camera view of a frame containing a person. Center: temporal signatures, $\overline{RGB}_k(t,h)$. Right: Bodyprints.

4.2 Matching Metric

In order to compare bodyprints we propose a normalized weighted correlation coefficient. Assume that we want to compare the j and k bodyprints. First, we compute a weight factor for each height that is related to the joint confidence on the elements of both bodyprints:

$$\mathcal{W}(h) = \min\left(\mathcal{C}_j(h), \mathcal{C}_k(h)\right) \tag{5}$$

The use of $\mathcal{W}(h)$ allows to compare bodyprints with missing values (due to occlusions), and to put more emphasis in the bodyprint portions that have been visible for more time.

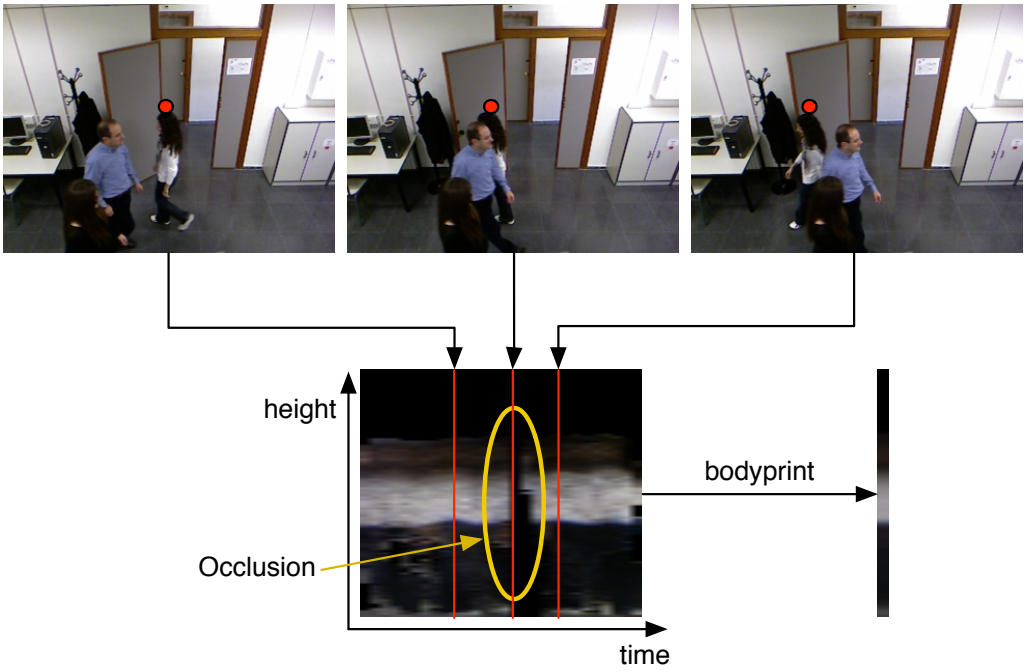

Figure 9: Example of track with temporary occlusion. Above: frames at different times. Below: temporal signature and bodyprint.

Next, we compute a weighted mean for each track which is used to compensate changes in brightness:

$$M_j = \left(\sum_h \mathcal{W}(h) \frac{\overline{R}_j(h) + \overline{G}_j(h) + \overline{B}_j(h)}{3} \right) / \left(\sum_h \mathcal{W}(h) \right) \tag{6}$$

where $\overline{R}_j(h)$, $\overline{G}_j(h)$ and $\overline{B}_j(h)$ are each of the color components of $\overline{RGB}_j(h)$. Many authors obtain different mean values for each color component and compensate each channel independently. However, we think that using a single mean value is useful to distinguish among colors (green vs. red for instance).

Finally, the normalized correlation coefficient is computed as:

$$\rho(j,k) = \frac{S_{jk}}{\sqrt{S_{jj}\, S_{kk}}} \tag{7}$$

$$\begin{aligned} S_{jk} &= \sum_h W(h)(\overline{R}_j(h) - M_j)\,(\overline{R}_k(h) - M_k) + \\ &+ \sum_h W(h)(\overline{G}_j(h) - M_j)\,(\overline{G}_k(h) - M_k) + \\ &+ \sum_h W(h)(\overline{B}_j(h) - M_j)\,(\overline{B}_k(h) - M_k) \end{aligned} \tag{8}$$

Notice that $\rho(j,k)$ is between -1 and 1. Maximum similarity corresponds to 1. The proposed measure of similarity is robust to:

- Differences in contrast.
- Differences in brightness.
- Differences in saturation

On the other hand it is sensitive to:

- Differences in colour scheme. (dark in the upper part and bright on the lower or viceversa)
- Differences in color (red vs. blue)
- For the same color scheme, for instance light in the upper part of the body and dark on the lower, differences of the height at which light becomes dark.

The proposed measure can even discriminate between people with similar clothing as long as their height is different because height information (through h) is implicit in the measure.

5 Results and discussion

This section presents the re-identification results that we have obtained in videos recorded using two kinect cameras. Each video contains more than 40 different people and the whole data set has been made publicly available (color, depth and groundtruth information) [2]. To our knowledge, this is the first public database for re-identification that includes depth information. Next subsection describes the dataset; Section 5.2 presents the evaluation methodology and the obtained results; Section 5.3 shows the system performance on more challenging scenarios, with different camera perspective and illumination.

5.1 Data set description

In this work, we have used two kinect cameras (camera 1 and camera 2) located approximately at the same location but pointing at opposite directions. Since people in the videos can approach or recede from the cameras, frontal and rear views of people were obtained.

We recorded four video sequences with each camera at a frame rate of approx. 15 images per second and image resolution of 320x240 pixels. In two of them, people are only approaching to the camera and in the other two, people are only receding. So that a total of eight sequences were obtained using both cameras. Each video is named using following the convention: C1F1 means camera 1, front view and pass 1; C2R1 means camera 2, rear view and pass 1 and so on. So that the eight combinations are: C1F1, C1F2, C1R1, C1R2, C2F1, C2F2, C2R1 and C2R2.

One problem that we have found when recording the sequences is that the automatic gain control of the RGB sensor produces large color variations. Sample videos that illustrate this point can be downloaded from [2].

5.2 Tests description and results

All the videos were processed and the corresponding bodyprints stored in separate databases. In order to evaluate the re-identification performance of the method presented in this paper, we conducted two experiments that are described next.

Experiment 1

In this experiment, people recorded by camera 1 is searched across some videos recorded by camera 2. In this experiment, it is assumed that query and target bodyprints are both from frontal or both from rear views. This situation is common in many surveillance scenarios. For instance, one camera captures people entering into a shop and another one captures people at the exit. In the experiment, frontal bodyprints from sequences C1F1 and C1F2 are used as query and then searched in each of the two target databases:

- C2F1
- C2F2

Also, every rear bodyprint from C1R1, C1R2 has been used as query in each of the following two databases:

Figure 10: Samples of the front-front matches found by the system with bodyprints obtained from cameras 1 and 2.

- C2R1
- C2R2

This makes a total of $40 \times 2 \times 2 \times 2 = 320$ different queries on databases of 40 people. The average correct re-identification performance that we have obtained is **93%**. Examples of correct matches found by the system are shown in figures 10 and 11 (for images with several people, the person matched is the one with an attached label on his/her head).

An example of incorrect match is shown in Fig. 12. In this case, the correct match had the second highest correlation coefficient and it was very similar to the highest (0.87345 and 0.87212).

Figure 11: Samples of the rear-rear matching results with bodyprints obtained from cameras 1 and 2.

Figure 12: Example of wrong match.

Experiment 2

In this experiment, people are re-identified using the same camera. The key difference compared to the previous experiment is that frontal and rear views are now compared. This situation is also common in two-way surveillance scenarios where only one camera is installed.

In this experiment, bodyprints from sequences C1F1, C1F2 are used as query and then searched in each of the two target databases:

- C1R1
- C1R2

Also, every bodyprint from sequences C2R1, C2R2 is searched in each of the following target databases:

- C2F1
- C2F2

This makes a total of $40 \times 2 \times 2 \times 2 = 320$ different queries on databases of size 40. The average correct re-identification obtained in this experiment drops to **55%**. Examples of correct and incorrect matches are shown in Fig. 13. The main causes of error in this experiment are the presence of logos on T-Shirts, backpacks, long hair in the case of some girls and opened jackets. All these situations greatly change the appearance between frontal and rear views. Examples of all these situations are depicted in Fig. 13.

5.3 Performance on different scenarios

In this section we are presenting results obtained in two additional scenarios that illustrate the robustness of the proposed approach.

The first one is an office environment with two cameras located at different heights in different locations with different illuminations. The number of people working in the office is 25. Examples of re-identifications can be seen in Fig. 14. Despite the variations in illumination, distance to the camera, and orientation with respect to the camera every person in this scenario was correctly re-identified. Probably

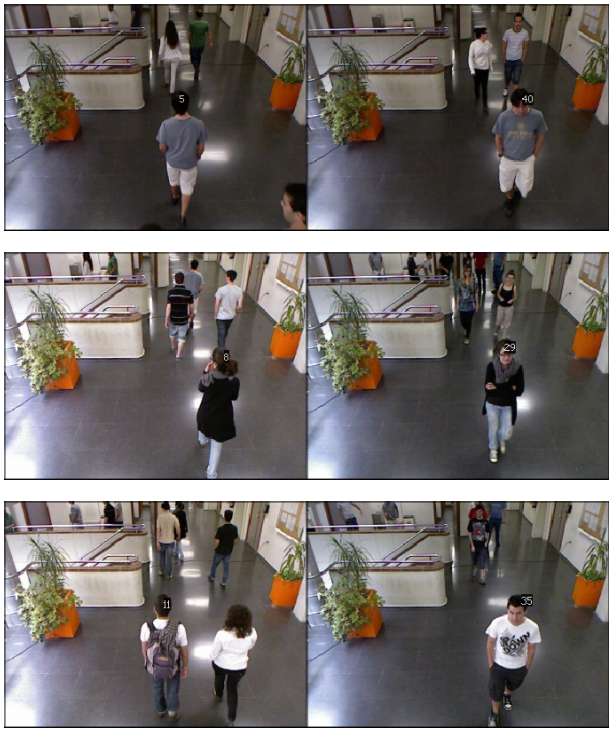

Figure 13: Example of mixed view matches. The two first ones are correct. The last one was incorrect due to the backpack that obviously created a very different bodyprint from the front and from the back.

the fact that people do not wear backpacks or open coats while working and the smaller number of people helped to achieve this result even in the presence of large illumination and point of view variations.

The second additional scenario was a small supermarket, where two cameras were placed at different locations: entrance, and exit.

In this case each person detected at the exit was compared against the people that entered the shop in the previous 30 minutes. The number of candidate people for each query was between 15 and 50 depending on the moment of the day. The total number of queries was over 200. A correct re-identification ratio above 90% was obtained. Figure 15 illustrates one example of correct re-identification in this scenario.

6 Conclusions

In this paper a method to match people using RGB-d videos has been presented. The obtained results are very promising using a medium size database. The size of this database is very representative for many re-identification scenarios, like retail.

Figure 14: Examples of re-identification in an office environment.

The method has proved to be robust against differences in illumination, point of view, and momentary partial occlusions. The errors are due mainly for the following reasons:

- Very similar appearance of two different people. (example Fig. 12)
- Very different appearance of the same person from the point of view of each camera. (people with open jackets, backpacks, etc.) (example Fig. 13-bottom)

In order to cope with the first problem more complex models can be used. These models would allow to be more discriminative among different people. Some of the models on which we are working are:

- Use both color mean and variance as a function of height in the bodyprints. Variance could be used

Figure 15: Example of re-identification in a supermarket environment. Left: entrance. Right: exit.

as a feature itself or as a confidence weight for comparing means. Variance based features would allow to differentiate between unimodal histograms and bimodal histograms with the same mean at each height.

- Features based on histograms of gradient orientations such as SIFT or SURF, could be used to recognize interest points (such as logos on T-shirts). Since the distance and orientation to the camera is known, the scale and rotation invariance could be disabled in these descriptors to be more distinctive.

To solve the second problem, in scenarios where matching between frontal and rear views is required, we are working on models that take into account not only the height of each point but its relative angular position with respect to the person axis. This way we could compare colors of points that belong approximately to the same location on the person such as the shoulders and sides of the person.

In our opinion, the key aspects of the performance of the method presented in this paper are:

- The high precision of depth measurements that allows to place each pixel in the right position in space.

- The use of all the information of one person in different times along the track. This makes the approach robust to occlusions.

- The matching metric, that is not sensitive to global changes in illumination and captures well differences in the pattern of intensity-color as a function of height.

Finally, since one of the problems in computer vision is to have common data sets for fair comparison, we have made publicly available the main dataset used in this work (the first one that includes depth

information to our knowledge). It should be pointed out that depth information can always be ignored so that existing approaches that only use RGB information can be compared too.

References

[1] http://www.pointclouds.org/.

[2] http://www.gpiv.upv.es/kinect_data/.

[3] http://www.primesense.com/.

[4] http://www.openni.org/.

[5] http://opencv.willowgarage.com/.

[6] http://www.ros.org/.

[7] http://kinectforwindows.org/.

[8] A. Angelo and J.-L. Dugelay. People re-identification in camera networks based on probabilistic color histograms. In *Visual Information Processing and Communication SPIE Electronic Imaging*, volume 7882, 2011.

[9] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pages 1413–1416, Washington, DC, USA, 2010.

[10] S. Bandk, E. Corvee, F. Br andmond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 435 –440, september 2010.

[11] S. Beucher and F. Meyer. *The Morphological Approach to Segmentation: The Watershed Transformation*, pages 433–481. Marcel-Dekker, 1992.

[12] N. Bird, O. Masoud, N. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):167 – 177, june 2005.

[13] I. Bouchrika, J. N. Carter, and M. S. Nixon. Recognizing people in non-intersecting camera views. In *International Conference on Imaging for Crime Detection and Prevention*, 2009.

[14] M. Buml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *International Conference on Advanced Video and Signal-Based Surveillance*, 2010.

[15] D. Chen, A. Bharucha, and H. Wactlar. People identification through ambient camera networks. In *International Conference on Multimedia and Ambient Intelligence*, 2007.

[16] I. de Oliveira and J. de Souza Pio. People reidentification in a camera network. In *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 461 –466, dec. 2009.

[17] H. Dee and S. Velastin. How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications*, 19:329–343, 2008.

[18] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Asian Conference in Computer Vision*, 2010.

[19] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: Problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–25, 2011.

[20] T. Gandhi and M. Trivedi. Panoramic appearance map (pam) for multi-camera based person re-identification. In *International Conference on Video and Signal Based Surveillance*, 2006.

[21] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007.

[22] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, pages 262–275, Berlin, Heidelberg, 2008.

[23] M. Grimaud. A new measure of contrast: the dynamics. In SPIE, editor, *Image Algebra and Morphological Image Processing III*, pages 292–305, 1992.

[24] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Second ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1 –6, sept. 2008.

[25] U. H. Office. i-LIDS multiple camera tracking scenario definition. 2008.

[26] C. Vachier and F. Meyer. Extinction value: a new measurement of persistence. In *IEEE Workshop on Nonlinear Signal and Image Processing*, 1995.

[27] W. von Hansen. Automatic detection of zenith direction in 3d point clouds of built-up areas. In *PIA07 - Photogrammetric Image Analysis*, pages 93–97, 2007.

[28] D.-J. Wang, C.-H. Chen, T.-Y. Chen, and C.-T. Lee. People recognition for entering and leaving a video surveillance area. In *Fourth International Conference on Innovative Computing, Information and Control*, pages 334 –337, dec. 2009.

[29] Z. Zhang and N. F. Troje. View-independent person identification from human gait. *Neurocomputing*, 69(1-3):250 – 256, 2005. Neural Networks in Signal Processing.

[30] T. Zhao, M. Aggarwal, R. Kumar, and H. Sawhney. Real-time wide area multi-camera stereo tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 976 – 983, june 2005.