UNIVERSITY OF PADOVA
DEPARTMENT OF INFORMATION ENGINEERING

MASTER'S THESIS

ALGORITHMS FOR PEOPLE RE-IDENTIFICATION FROM RGB-D VIDEOS EXPLOITING SKELETAL INFORMATION

Deniz Tartaro Dizmen

SUPERVISOR: Emanuele Menegatti

ASSISTANT SUPERVISORS: Matteo Munaro, Stefano Ghidoni

Accademic Year 2012–2013

Alla mia famiglia

Part of the inhumanity of the computer is that,
once it is competently programmed and
working smoothly, it is completely honest.

-Isaac Asimov-

Abstract

In this thesis, a novel methodology to face the people re-identification problem is proposed. Re-identification is a complex research topic in Computer Vision representing a fundamental issue, especially for intelligent video surveillance applications. Its goal is to determine the occurrences of the same person in different video sequences or images, usually by choosing from a high number of candidates within a dataset. In our method, a highly distinctive and compact feature-based signature is generated for each person by exploiting the skeleton provided by a consumer RGB-D sensor such as Microsoft Kinect. This signature is created by concatenating in a specific order the local descriptors extracted around the joints of the human body. We tested and compared a number of state of the art 2D and 3D feature descriptors on two public datasets for people re-identification with RGB-D sensors. Our approach achieves very good results in terms of both recognition accuracy and framerate with respect to standard methods which exploit SIFT keypoint detector or color histograms.

Sommario

La tesi propone una nuova metodologia per affrontare il problema della re-identificazione di persone, un argomento di ricerca che si inserisce all'interno della Computer Vision e che trova applicazione principalmente nei sistemi di videosorveglianza. L'obiettivo del problema accennato è quello di determinare quando la stessa persona è presente all'interno di diverse sequenze video, considerando che durante le riprese sono state acquisite le immagini di un elevato numero di soggetti. Nel metodo proposto in questo lavoro, un descrittore globale della persona, compatto ed altamente informativo, viene generato sfruttando le informazioni della posa e della posizione dello scheletro ottenute da un sensore RGB-D come Microsoft Kinect. Questo descrittore globale viene costruito concatenando in un ordine ben preciso ogni descrittore locale che codifica l'informazione estratta dai punti chiave del corpo, ossia i giunti scheletrici. Sono stati testati e confrontati un vasto numero di tecniche per descrivere una persona sia attraverso i dati 2D che 3D, usando due dataset pubblici adatti per la re-identificazione di persone, poiché sono disponibili i valori RGB-D per ogni immagine. L'approccio qui proposto ottiene risultati eccellenti sia in termini di accuratezza che di tempi di esecuzione, se paragonato ai metodi classici allo stato dell'arte che sfruttano i keypoint SIFT o si servono degli istogrammi di colore.

Contents

Chapter 1

Introduction

The well known first Gordon Moore's law, written in 1965, states that the number of transistors on a chip doubles approximately every two years. Assuming that it is true, present day computers are about two billion times more powerful than in the '60s.

The exponential growth in terms of computational power has facilitated the birth of several new research areas such as *Computer Vision* (CV), which is a branch of robotics and intelligent systems. CV aims to simulate what I believe to be the most amazing human sense: the sight. Recent studies suggest that our eyes can transmit over the optic nerve an amount of data equal to ten million of bits per second[1]. However, the complexity of our visual system goes far beyond the task of seeing only: we are able to recognize objects around us and give them their common nouns, to predict where a person will move just a moment later, to perceive risks, and so on, with our brain that processes external inputs combining them with our life experiences.

Nowadays, the broader field of application of robotics is the industrial automation, in which CV is involved in the computerized visual inspection for quality control. Other sectors concern autonomous vehicles (e.g. Nasa's *Curiosity* employed on Mars), intelligent video surveillance systems, etc. In all the mentioned applications, the principal tasks of CV can be summarized with:

- *Recognition*: one or several pre-specified or learned objects or object classes can be recognized, usually together with their 2D positions in the image or 3D poses in the scene;

- *Identification*: an individual instance of an object is recognized. Examples include identification of a specific person's face or fingerprint, or identification of a specific vehicle;

- *Detection*: the image data are scanned for a specific condition. Examples include detection of possible abnormal cells or tissues in medical images or detection of a vehicle in an automatic road toll system;

- *Reconstruction*: given one or (typically) more images of a scene, or a video, scene reconstruction aims at computing a 3D model of the scene.

This thesis faces one of the most recent CV's problems: people re-identification (re-id). It is a complex research topic representing a fundamental issue especially for intelligent video surveillance and ambient intelligence applications. Its goal is to determine the occurrences of the same person in different video sequences or images, usually by choosing from a high number of candidates within a dataset. The main difficulties of this task can be identified in perspective changes, illumination variance, occlusions and the considerable number of individuals having similar appearance. Several different approaches to solve re-id problem have been tested in literature; the most important and recent are briefly introduced in the next section.

1.1 Related work

The main approaches in literature for re-id can be divided into three categories: i) appearance based, ii) appearance and shape based, iii) shape based. The first category includes the majority of state of the art works and aims at describing people by extracting information from chromatic and texture traits [2, 3, 4, 5, 6, 7]. Recent works belonging to the second category [8, 9, 10] exploited both appearance and shape to increase the total system accuracy using RGB-D sensors such as Microsoft Kinect, available at increasingly lower prices. The last approach is based on features extracted from body shape only [11].

Yoon et al [2] model the appearance by exploiting color and spatial information and suggesting a descriptor named color/path-length (CPL). The pathlength of a pixel is defined as the normalized length of the shortest path from the top of the head to the pixel inside the silhouette. By normalizing, pathlength can be used as a scale-invariant feature. In addition, color information is used defining Brightness = R + G + B, where R, G and B are the values of the color channels, and three color ratios, red = R/Brightness, green = G/Brightness and blue = B/Brightness. The Kullback-Leiber distance, presented for the first time in [12], is used to evaluate the similarity between two persons.

Similarly, Cong et al [3] propose a feature called color-position histogram. More specifically, the silhouette is vertically divided into N equal parts and each of them is characterized by its mean color. The color-position histogram is composed of 3 N values (while working with three color channels). A normalization of the color channels is performed to be invariant to lighting conditions. The final algorithm is based on a SVM in the matching step.

Hu et al [4] use SIFT for local description and Correlograms for global description of the individuals appearance. A correlogram is a graph in which autocorrelation coefficients are plotted against time. The similarity measure of a person in the training frame to another person in the queried frame is computed as the product of color and feature similarities. The global and local descriptors are used for training a strong classifier on-line with Adaboost to distinguish a newly detected person as tracked or new occurrence.

Farenzena et al [5] extract features for three complementary appearance aspects: color information, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy. These features are weighted

by the distance with respect to the vertical axis of the human body, so that the effects of pose variations are minimized. They called their method SDALF and introduced a new similarity measure that combines all the extracted features; the SDALF total distance is the sum of three coefficients: i) histogram distance, ii) Recurrent High-Structured Patches (RHSP) distance and iii) Maximally Stable Color Regions (MSCR) distance. The first two distances i) and ii) are calculated by means of the Bhattacharyya distance [13] and iii) is calculated by means of the Euclidean distance of the regions centroid and their main color.
An alternative approach is explained in [6], in which Jungling et al use the SURF feature detector and descriptor over thermal images. They motivate the choice of working on thermal images because the variation in person appearance is rather limited compared to visible wavelength imagery.
De Oliveira and De Souza [7], similarly to other state of the art techniques, combines color information with features. Unlike [2, 3], the color information is not given by a traditional normalized color histogram, but by the Hue histogram, which is invariant to brightness and Gamma. In all frames, N interest points are detected with SURF. Their final signature is calculated by concatenating the Hue values with the SURF feature descriptor.

The second category of works faces the re-identification topic by exploiting both appearance and shape information. Baltieri et al [8] propose an interesting framework: first, a people detection module analyzes videos of all the cameras within a camera network. All frames of all the views are combined in order to detect people and estimate their position in each frame. A short-term tracking system is exploited to locally match the detections using geometrical information and spatial constraints only. A body model of each person is then created and a long term tracking matches and merges together the trajectories that are recognized to belong to the same person. The matching stage aims at finding correspondences between pairs of models using shape and color information in order to calculate their similarity.
A combination of multi-attribute properties is performed by Liu et al [9] by extracting three different features from individuals by using a RGB-D sensor: biometrical, appearance and motion attributes. The biometrical characteristics are the people height and their shoulder breadth while the appearance features are based on people skin color, their clothing color and texture. The person is divided into 23 blocks covering skin, coat, trousers and luggage and each part is described with a HSV color histogram to be robust against illumination changes. The motion attributes include squatting, running and wandering. Their experimental results confirm that, by using all extracted attributes, the system precision is higher with respect to using appearance based features only.
Oliver et al [10] face the re-identification problem in an interesting manner. They create a descriptor based on a 3D cylindrical grid that stores color variations with angle and height. The spatial coordinates of a points of a particular person are changed to cylindrical coordinates centered at the center of mass, and angularly aligned so that $\theta = 0$ corresponds to the walking direction. The cylindrical grid is organized with N angular bins and N_h height bins in which the points of a person are assigned. The mean color of the points falling into each cylindrical bin and the number of points w that they contain are computed and the descriptor vector with

chromatic information and w is created. It summarizes all the information about the appearance of a person along time as a function of angle and height. In the matching phase, the distance metric is a sum of the normalized chromatic feature distances weighted by a coefficient that includes w.

Finally, a shape-based re-identification is performed by Barbosa et al [11]. The strength of this method is that, unlike all works reported above, no chromatic information is used, thus a person can be recognized also with different clothes. The targeted scenario is a long term re-identification. They extract 3D soft-biometric cues directly from range data given by a RGB-D sensor. Two types of features are computed: skeleton-based and surface-based. The former are based on the combination of distances between joints and distances of joints to the ground plane, the latter are composed of some geodesic distances on the mesh surface computed between different joints pairs. These latest measures give an indication of the curvature and of the size of specific body regions. During the matching stage, the squared distance between each feature has been used as a similarity metric to compare a person in the training set to another person in the testing set.

1.2 Our approach

We propose a novel approach with respect to the state of the art techniques, which widely exploit color histograms or features extracted from keypoints obtained by a detection module. Our method instead extracts features from special keypoints: the joints of the human body provided by a skeletal tracking algorithm. The advantages of using these keypoints are mainly two: i) their positions are spread all over the human body; ii) computational complexity of the system is reduced. With regards to i), keypoints are generally obtained as output of a detection phase, but we can not be assured that they are detected in descriptive regions. Our keypoints, instead, are located in very particular positions given directly from the skeletal tracker developed by Microsoft or by OpenNI framework. Moreover, since we already know the correspondences between pairs of keypoints in different frames, we do not need to perform a keypoint matching phase: instead, we directly calculate the distance between final person signatures obtained by concatenating in a specific order each local descriptor extracted from the skeleton keypoints. Distance computation is fast because we manage a small number of keypoints; then, time complexity is low since we do not need to detect and match keypoints. To prove these benefits, we compared the results of our approach to those obtained when detecting keypoints in a traditional way and we showed that our method performs better, while being both simple and fast.

1.3 Thesis structure

The thesis is organized as follows: Chapter 2 explains the theory behind recent feature descriptor algorithms used in this work. Details of the method we propose are explained in Chapter 3. Experimental results and conclusions are presented in Chapters 4 and 5.

Chapter 2

Local descriptors

In Chapter 1, we mentioned that our re-identification approach is based on feature extraction and classification. In computer vision, feature detection and extraction target to identify interesting points, called keypoints, such as a corner or an edge, and mathematically describe them. An ideally perfect feature is highly discriminative to always distinguish a particular keypoint from another one within all images of the same object. It is important to notice that the feature extraction process can be logically split into keypoint localization and keypoint description. In this work, we do not make use of traditional keypoint detection algorithms, thus this chapter only deals with the description algorithms.

In the last decades, a lot of descriptors were implemented and many of them are better than the others in some cases. Thus, we studied and compared some of them in order to understand which ones are more suitable for our approach. We considered both 2D and 3D features: the first type extracts information from images and the second one from 3D data, such as a point cloud. All the descriptors we tested are open source and available in OpenCV[1] 2.4.2 (2D descriptors) and PCL[2] 1.7 (3D descriptors). In this section, we will give an overview of the descriptors we tested.

2.1 2D features

O*pen* ***S****ource* ***C****omputer* ***V****ision*, OpenCV, is probably the most advanced and supported library for computer vision tasks. We worked with the latest available release (2.4.2) at the moment of writing this thesis. The 2D features we considered can be split into two categories based on the values contained in the descriptor vector: real or binary. The first category includes SIFT and SURF while BRIEF, ORB and FREAK are examples of binary descriptors. In addition to the descriptor type, another difference between the real and binary descriptors concerns the distance that should be used for matching them: Euclidean distance for real values, Hamming distance for bitstrings[14].

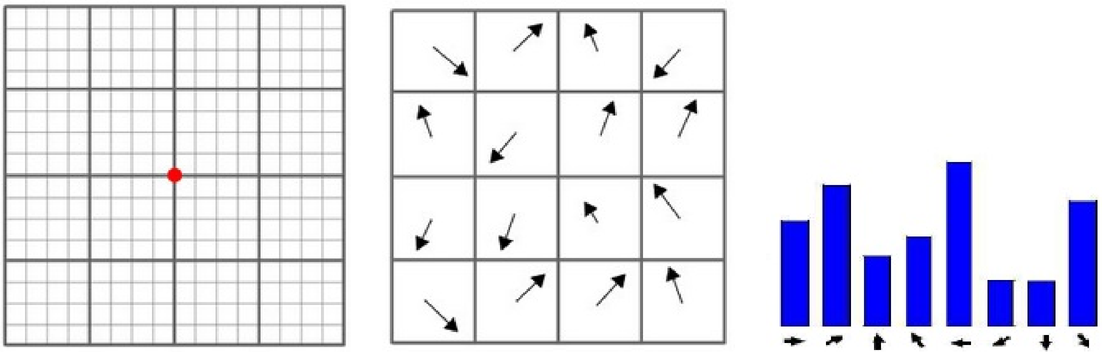
[1] `http://opencv.org`.

[2] `http://pointclouds.org`.

2.1.1 SIFT

***S**cale-**I**nvariant **F**eature **T**ransform*, SIFT, is a keypoint detector and feature descriptor presented by Lowe in 1999. Details of the algorithm can be found in [15]. The SIFT features are invariant to image scale and rotation and also robust to changes in illumination, noise, and minor affine transformations.

SIFT descriptor is an array of 128 floats. A 16 × 16 window is centered on the keypoint. This windows is broken into sixteen 4 × 4 windows. An orientation histogram, with 8 bins, is computed from magnitude and orientation values within each 4 × 4 windows, as in Figure 2.1c.



(a) 16 × 16 window around the keypoint.

(b) Gradient magnitude and orientations.

(c) 8-binned orientations histogram.

Figure 2.1: SIFT feature descriptor creation.

The magnitudes depend on the distance from the keypoint; then, gradients are weighted by a Gaussian function. The descriptor becomes a vector of all the values of these histograms. Since there are 4 × 4 = 16 histograms each with 8 bins, the vector has 128 elements. This vector is then normalized to unit length. To reduce the effects of non-linear illumination a threshold of 0.2 is applied and the vector is again normalized (see Figure 2.2).

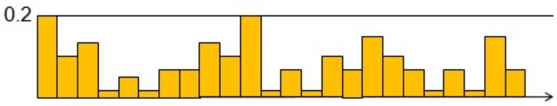


Figure 2.2: Threshold of 0.2 applied to the SIFT descriptor vector.

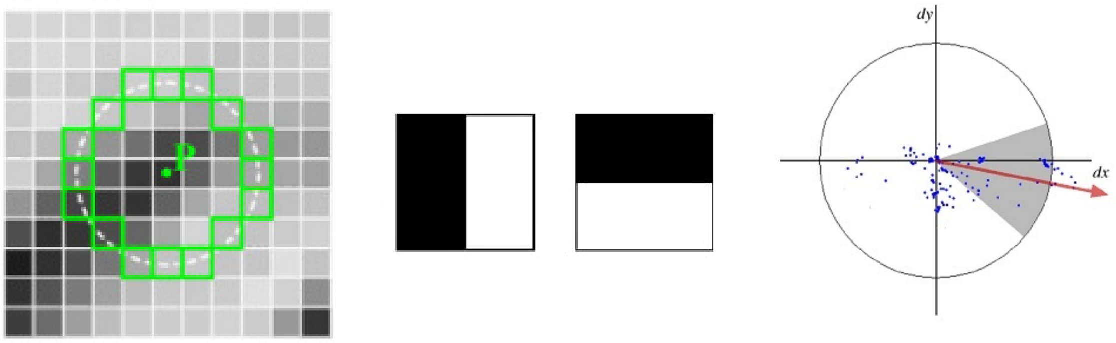
SIFT reaches a very good performance in the most cases, but it is very slow with respect to the other descriptors. One of the causes is because SIFT detects so many keypoints and finds so many matches. In our case the number of the keypoints is fixed and equal to the number of tracked skeletal joints. So, this is not a problem for us.

In 2004, a variant of SIFT named PCA-SIFT[16] was developed by Yan Ke *et al* to reduce the descriptor size. Here, the descriptor is a vector of image gradients in x and y direction computed on a patch centered on the keypoint. The gradient region is sampled at 39×39 locations, therefore the vector is of dimension $39 \times 39 \times 2 = 3042$.

The dimension is reduced to 36 elements with PCA[17]. A comparison between SIFT and PCA-SIFT[18] shows that the former outperforms the latter in scale, rotation and blur changes while PCA-SIFT is better in illumination and less time in keypoint detection and descriptor extraction is needed.

2.1.2 SURF

SURF, abbreviation of ***S****peeded-****U****p* ***R****obust* ***F****eature*, was developed by Bay et al[19] in 2006. It is a keypoint detector and descriptor invariant to image scale, illumination changes and orientation, partially inspired by Sift descriptor.



(a) Circular region around SURF keypoint.

(b) Haar wavelet examples.

(c) Dominant SURF keypoint direction.

Figure 2.3: SURF keypoint orientation assignment.

To compute the descriptor vector, a $20\,s \times 20\,s$ window, s = scale at which the keypoint is detected, is placed on the keypoint, accordingly with its orientation (Figure 2.3c). This region is split into a 4×4 sub-regions. For each of them the Haar-wavelet (Figure 2.3b) responses in horizontal, d_x, and in vertical, d_y, directions are computed and weighted with a Gaussian kernel ($\sigma = 3.3\,s$). The responses d_x and d_y are summed within each sub-region. Furthermore, also the absolute values $|d_x|$ and $|d_y|$ of the responses are summed. Then, the final descriptor is a vector $\mathbf{v} = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$. The vector $\mathbf{v}$ is normalized to unit length. The final descriptor size is then $4 \times 4 \times 4 = 64$ floats (number of sub-regions × number of vector components) as in Figure 2.4.

An alternative version of the descriptor, SURF-128, doubles some informations to be more distinctive. The sums of d_x and $|d_x|$ are computed separately for $d_y < 0$ and $d_y > 0$. Similarly, the sums of d_y and $|d_y|$ are done according to the sign of d_x.

SURF is faster than SIFT. This is mainly due to the keypoint detection algorithm. In SURF, Hessian-Laplacian is used to approximate Laplacian of Gaussian; SIFT, instead, uses difference of Gaussian. The same result emerges in [18], where SURF outperforms SIFT in execution time and illumination changes only, achieving good results in image scale, blurring and affine transformations.

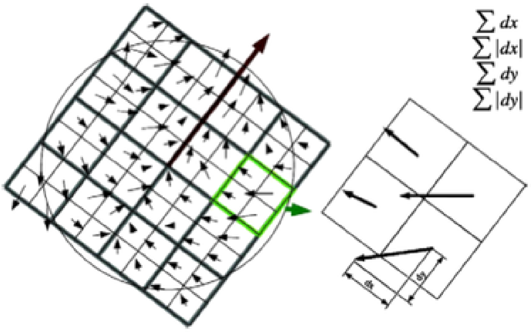


Figure 2.4: Sums in d_x and d_y directions for SURF descriptor.

2.1.3 BRIEF

Presented in 2010 by Calonder et al, BRIEF[20] (***B**inary **R**obust **I**ndependent **E**lementary **F**eatures*) is a feature descriptor algorithm. Unlike SIFT and SURF, the descriptor is composed by a binary string. Each bit of the vector is the result of an intensity test on a pair of a particular patch centered on the keypoint. More specifically, the intensity test is:

$$\tau(\mathbf{p}; x, y) := \begin{cases} 1 & \text{if } \mathbf{p}(\mathbf{x}) < \mathbf{p}(\mathbf{y}) \\ 0 & \text{otherwise} \end{cases}, \tag{2.1}$$

where $\mathbf{p}$ is the patch and $\mathbf{p}(\mathbf{x})$ is the intensity of the pixel $\mathbf{x} = (u, v)^{\mathrm{T}}$ in $\mathbf{p}$. A number n_d of $(\mathbf{x}, \mathbf{y})$ pixel pairs defines a set of binary tests. Therefore, the final descriptor is the following n_d-dimensional binary string:

$$f_{n_d}(\mathbf{p}) := \sum_{1 \le i \le n_d} 2^{i-1} \, \tau(\mathbf{p}; x, y) \, . \tag{2.2}$$

In their paper, Calonder et al consider $n_d = 128, 256, 512$. An important choice involves the spatial arrangement of the $(\mathbf{x}, \mathbf{y})$ pairs in equation 2.1. They studied five different patches:

I. $(\mathbf{X}, \mathbf{Y}) \sim$ i.i.d Uniform$(-\frac{S}{2}, \frac{S}{2})$

II. $(\mathbf{X}, \mathbf{Y}) \sim$ i.i.d. Gaussian$(0, \frac{1}{25}S^2)$

III. $\mathbf{X} \sim$ i.i.d. Gaussian$(0, \frac{1}{25}S^2)$, $\mathbf{Y} \sim$ i.i.d. Gaussian$(X_i, \frac{1}{100}S^2)$

IV. The $(\mathbf{x_i}, \mathbf{y_i})$ are randomly sampled from discrete locations of a coarse polar grid introducing a spatial quantization

V. $\forall i : \mathbf{x_i} = (0, 0)^{\mathrm{T}}$ and $\mathbf{y_i}$ takes all possible values on a coarse polar grid containing n_d points

Figure 2.5 shows graphically the configurations explained above.

The small dimension of the descriptor and the matching algorithm based on the Hamming distance, ensures a very high performance in terms of computational time. Brief suffers in image scale changing and in rotations, but it is not less robust than SIFT and SURF in image blurring and illumination changes, as shown in [20].

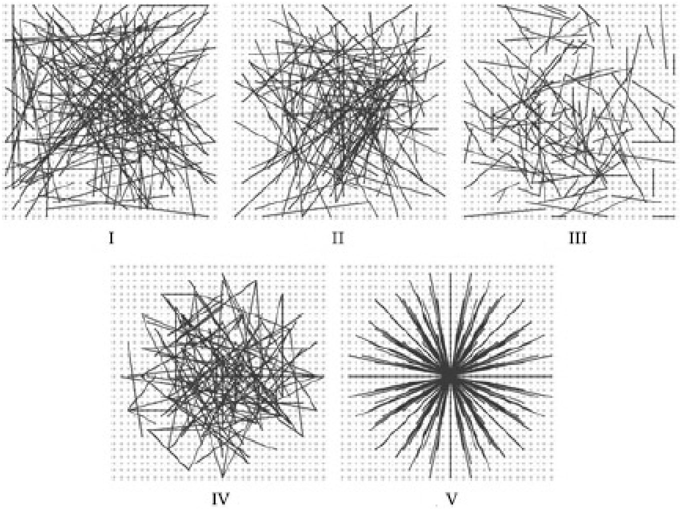


Figure 2.5: BRIEF patches. Images from paper[20].

2.1.4 ORB

Since BRIEF is not invariant to image rotation, Rublee et al in 2011 suggest a method to overcome this drawback. ORB[21], an acronym of ***OR****iented* ***BRIEF***, is the evolution of Calonder's algorithm. To add the rotation invariance, the artifice is to describe a keypoint according to its orientation. For any feature set of n binary tests at location (x_i, y_i), the $2 \times n$ matrix is defined:

$$\mathbf{S} = \begin{pmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{pmatrix} . \tag{2.3}$$

Then, $\mathbf{S}_\theta = \mathbf{R}_\theta \mathbf{S}$ is computed, where $\mathbf{R}_\theta$ is the rotation matrix that corresponds to the orientation θ of the patch. The Brief descriptor becomes:

$$g_n(\mathbf{p}, \theta) := f_n(\mathbf{p}) \mid (\mathbf{x}_i, \mathbf{y}_i) \in \mathbf{S}_\theta , \tag{2.4}$$

where $f_n(\mathbf{p})$ is the feature descriptor defined in equation 2.2. The angles increments are discretized of $2\pi/30$ radians.

As written in [21], ORB is very good in Gaussian blurring. Compared to SIFT and SURF, it is faster and reaches very good performances in illumination, rotation and scaling.

2.1.5 FREAK

F*ast* ***RE****tin****A*** ***K****eypoint* is another keypoint detector and descriptor extractor having a binary string descriptor vector. Introduced by Alahi et al[22] in 2012, FREAK computes the descriptor by comparing pixel intensities over a retinal sampling pattern. It is organized to simulate the saccadic search of the human eyes. The retinal sampling grid is circular and has a higher density of points and a minor gaussian smoothing near the center of the pattern, as shown in Figure 2.6.

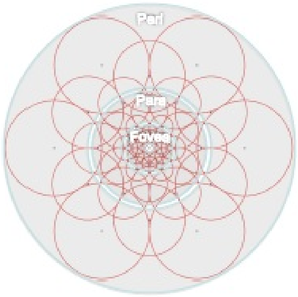


Figure 2.6: FREAK retinal sampling pattern. Images from paper[22].

Similarly to BRIEF and ORB, an intensity test is defined as:

$$T(P_a) = \begin{cases} 1 & \text{if } I(P_a^{r_1}) - I(P_a^{r_2}) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.5}$$

where $I(P_a^{r_1})$ is the intensity of the first perceptive field of the pair P_a. The descriptor is then created as follows:

$$F = \sum_{0 \leq a < N} 2^a \, T(P_a) \,, \tag{2.6}$$

in which P_a is a pair of perceptive fields, N is the size of the descriptor. In their experiments, Alahi et al observed that $N = 512$ is an ideal choice; adding more pairs the performance does not increase. The N pairs are distributed over the pattern according to the retinal conformation, such as in Figure 2.7.

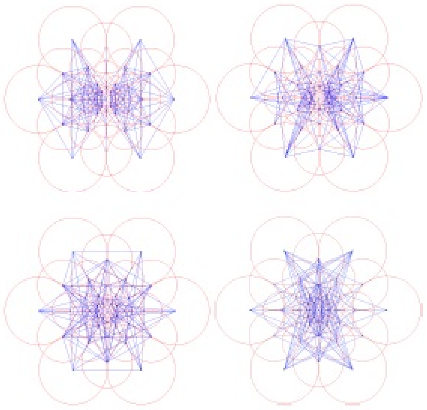


Figure 2.7: FREAK intensity tests. Images from paper[22].

In [22], it is asserted that FREAK reaches a good performance in terms of illumination, scale, affine transformations, blurring and execution time.

2.2 3D features

The 3D features we considered are all computed on a point cloud, which is a set of points in a N-dimensional space. Each point of a cloud contains several information, usually the XYZ-coordinates at least; moreover color channels can be added. The 3D feature description process is similar to the two-dimensional case, but descriptors are computed based on a mathematical characterization of 3D-keypoints instead of 2D ones. Typically, 3D features focus on the spatial arrangement of the points, i.e. on the point cloud shape. In addition, some descriptors characterize also the color information, making the descriptor more discriminative.

2.2.1 PFH

P*oint* ***F****eature* ***H****istograms*, PFH, is a keypoint descriptor introduced by Rusu et al[23], invariant to 6D pose of the surface. It aims to characterize a point based on its k-nearest neighbors. The relationships between the points and their estimated surface normals are represented as an histogram, as in Figure 2.8.

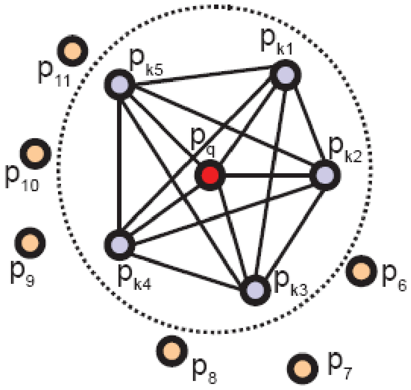


Figure 2.8: PFH keypoint in red and its k-neighbors in blue. Image from http://pointclouds.org/documentation/tutorials/pfh_estimation.php

Here, the point of interest and its neighbors within a sphere with radius r are fully interconnected in a mesh. Since k is the number of keypoint nieghbors, there are $O(k^2)$ connections in this mesh. Then, the total computational complexity of the algorithm for n keypoints is $O(nk^2)$. Consider two generic points, p_t and p_s and their normals n_t and n_s. A coordinate system fixed at one of the points is constructed, such as in Figure 2.9.

$$\begin{cases} \mathbf{u} = n_s \\ \mathbf{v} = \mathbf{u} \times \dfrac{(p_t - p_s)}{\|p_t - p_s\|_2} \\ \mathbf{w} = \mathbf{u} \times \mathbf{v} \end{cases} . \tag{2.7}$$

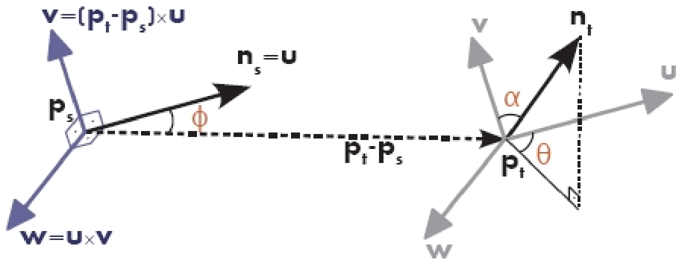


Figure 2.9: Coordinate system on a point. Image from `http://pointclouds.org/documentation/tutorials/pfh_estimation.php`

Then, the following angular features are computed:

$$\begin{cases} \alpha = \mathbf{v} \cdot n_t \\ \phi = \mathbf{u} \cdot \dfrac{(p_t - p_s)}{\|p_t - p_s\|_2} \\ \theta = \arctan(\mathbf{w} \cdot n_t, \mathbf{u} \cdot n_t) \end{cases} . \tag{2.8}$$

The quadruplet $< \alpha, \phi, \theta, d >$, where d is the euclidean distance between the two points, is computed for each pair in the interconnected mesh. The final PFH descriptor for a keypoint p_i, is created by binning all quadruplets into a histogram with b^4 bins. In the practical implementation of the algorithm, the fourth element of the quadruplet is not considered. In fact, the descriptor is a vector of $b^3 = 125$ floating elements, since in PCL the default implementation uses 5 binned histogram.

PFH is dependent on the quality of the surface normal estimations at each point. This requisite should be a problem with moving and deformable objects, like people, because the normals of the points at same location can change randomly after some time.

2.2.2 PFHRGB

As suggested by the name, PFHRGB adds color information to the standard PFH descriptor. Thus, the PFHRGB descriptor is a vector of 250 floats: the first 125 are obtained as explained in section 2.2.1, the last one is a 5 binned histogram of the color information around the keypoint, considering the same situation of Figure 2.8. It should be noticed that, characterizing a generic point with also its chromatic information, increases the robustness of the algorithm, but more time and memory are required.

2.2.3 FPFH

Developed by Rusu et al[24] in 2009, FPFH means ***Fast PFH***. It is another variant of the original PFH (section 2.2.1). The goal of this descriptor is to increase the performance in terms of computational complexity and memory consumption, without loss of robustness. The simplification is done by cutting some interconnections of the original mesh (Figure 2.8). It is a two step simplification: first, the so called Simplified Point Feature Histogram is computed for each keypoint p_q, based on a set of tuples $< \alpha, \phi, \theta >$ between itself and its neighbors as described in equation 2.8. The neighbours representation of FPFH is explained in Figure 2.10.

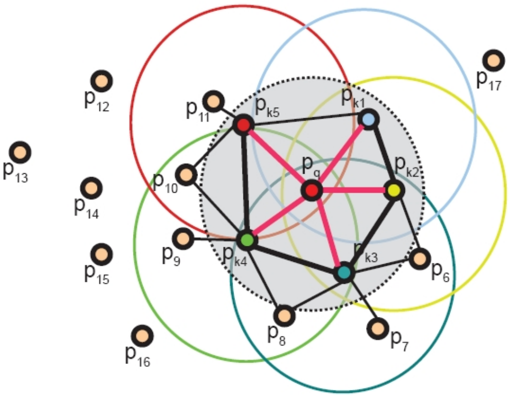


Figure 2.10: FPFH interconnections. Image from `http://pointclouds.org/documentation/tutorials/fpfh_estimation.php`.

In the second step, for each point, the k-neighbors are determined again, computing their SPFH. The final descriptor is created weighting all these SPFHs:

$$FPFH(p_q) = SPFH(p_q) + \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\omega_i}\cdot SPFH(p_i)\,, \tag{2.9}$$

where ω_i is a weight which represents the distance between p_q and p_i. Since FPFH does not fully interconnects all neighbors of a keypoint p_q, the total computational complexity for n keypoints is decreased from $O(nk^2)$ to $O(nk)$. In PCL, the default implementation uses a 11 binned histogram for each angular feature. These are separately calculated and then concatenated in the final descriptor into a vector with 33 floating entries.

2.2.4 SHOT

In 2010, Tombari et al proposed a novel local 3D descriptor and they called it SHOT[25] (***S****ignature of* ***H****istograms of* ***O****rien****T****ations*). As they wrote, before developing SHOT descriptor, they studied upon the impact of local 3D Reference Frames (RF) for 3D descriptors, which is a coordinate system centered in the point of interest. They observed that most of the existent RF cannot be uniquely selected and then they were able to define a unique and unambiguous RF. In their studies, Tombari et al were inspired by SIFT-like descriptors: the calculation of a RF is similar to SIFT or SURF keypoint orientation assignment. These 2D descriptors place a regular grid around the keypoint and describe the region of interest with an histogram based on the intensity gradients.

To emulate them, SHOT defines an isotropic spherical grid centered in the keypoint, such as in Figure 2.11.

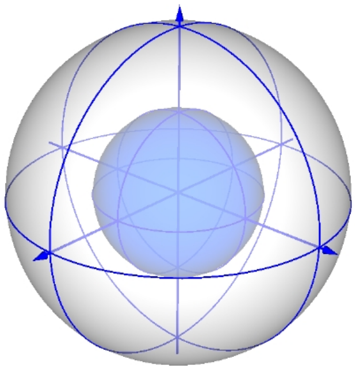


Figure 2.11: SHOT spherical grid. Image from paper[25].

For each sector of the grid an histogram of normals is computed. The final descriptor is formed by grouping all these histograms in a vector and normalizing it to unit length. Since the spherical grid is divided in 32 parts (8 azimuth divisions, 2 elevation divisions and 2 radial divisions) and each histogram has 11 bins, the SHOT descriptor has $32 \times 11 = 352$ values.

2.2.5 SHOTRGB

The year after SHOT (section 2.2.4), Tombari et al enriched SHOT descriptor adding it chromatic information. Thus, in 2011 they defined a novel descriptor, dubbed CSHOT[26] (***C****olor* ***SHOT***) or SHOTRGB.

The final descriptor is the concatenation of two histograms, based on the points shape and on color channels within a spherical grid around the keypoint (Figure 2.12).

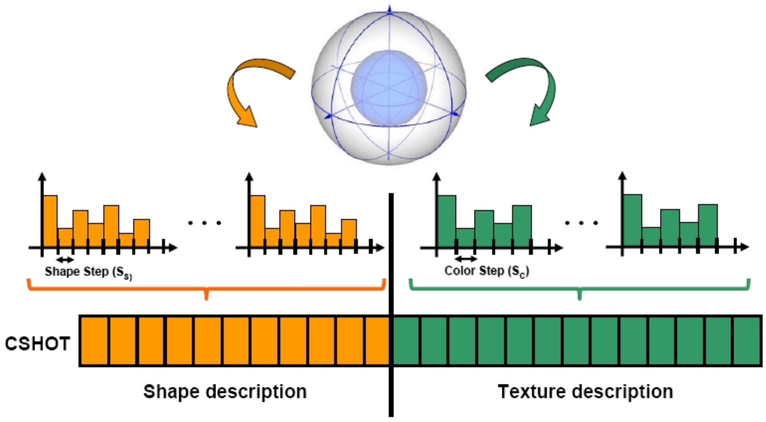


Figure 2.12: CSHOT spherical grid and its descriptor vector. Image from paper[26].

This descriptor is normalized to unit length such as SHOT. The color part of the descriptor is obtained by binning in a histogram the color information, after choosing a proper color representation and a comparison metric. In [26], RGB and CIELab color spaces were tested. Moreover, two different distance metrics are examined: dot product between RGB_P and RGB_Q, where P and Q are two points and the L_1 norm between the RGB triplet of the points P and Q, which is the sum of the absolute differences between the triplets RGB_P and RGB_Q. As a conclusion of their paper, Tombari et al wrote that the best combination between color space and distance metric is L_1 norm in CIELab color space. In PCL, the default implementation for SHOTRGB provides a descriptor vector with 1344 elements.

Chapter 3

Methodology

Our re-identification approach relies on computing a person descriptor by concatenating features at the body joints locations. Since we tested both 2D and 3D features, we exploited the Microsoft Kinect sensor for acquiring both RGB and depth information. Our method is logically split into three steps, which can be briefly summarized as follows:

1. *Recording*: using Kinect sensor we record some videos for each person in order to create a dataset for re-identification with RGB-D information, in which RGB, depth images and also body joint positions for each frame are stored.

2. *Skeleton-based Person Signature*: we describe each person within our datasets by extracting state of the art local descriptors (Chapter 2). We tested separately both 2D and 3D descriptors (on images and on point clouds). A compact signature is computed for all individuals in our database with a method explained in Section 3.2.

3. *Signature Matching*: each person in the testing set is matched with the most similar one in the training set. The similarity metric is the total distance between person signatures.

3.1 Recording

RGB-D sensors are becoming available at increasingly lower prices. This fact helped to their massive use in several applications, such as in the gaming industry and in computer vision. The Microsoft Kinect RGB-D sensor, shown in Figure 3.1, is composed by a standard RGB camera and an infrared projector-sensor pair.

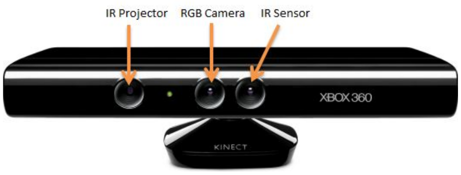
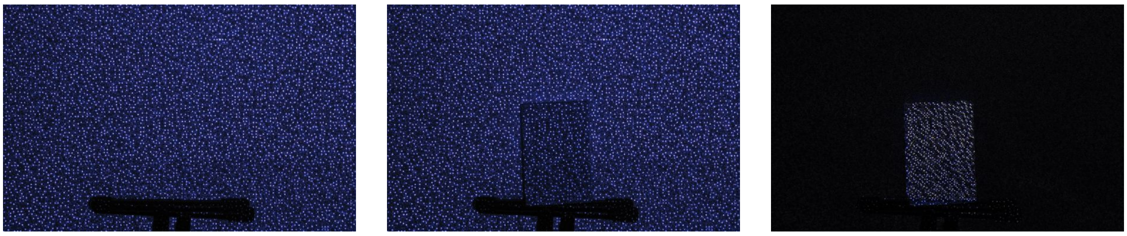


Figure 3.1: Components of the Kinect sensor: ir projector-sensor and RGB camera.

Exploiting the infrared projector-sensor pair, the Kinect can operate a 3D reconstruction of the scene[1]. For this purpose, a particular pattern is projected over the scene and by observing how it deforms due to obstacles and how it differs from a reference pattern, the Kinect detects the surfaces of the objects and can calculate the depth (i.e. the Z-coordinate) for each pixel in the RGB image. The Figure 3.2 shows an example in which the surface of a book is estimated by using the strategy described above.



(a) Reference pattern **(b)** Deformed pattern **(c)** Region of deformation

Figure 3.2: Kinect patterns and the region of deformation. Images from `http://www.futurepicture.org/?p=116`.

Another important feature we used in our approach concerns the skeletal tracking algorithm: we tested both Microsoft's tracker[27] and OpenNI's tracker[2]. These trackers provide us positions and orientations in 3D of the twenty joints of the human body for what concerns Microsoft and fifteen body joints for OpenNI. The names and positions of the joints for the two skeletal trackers are shown in Figure 3.3.

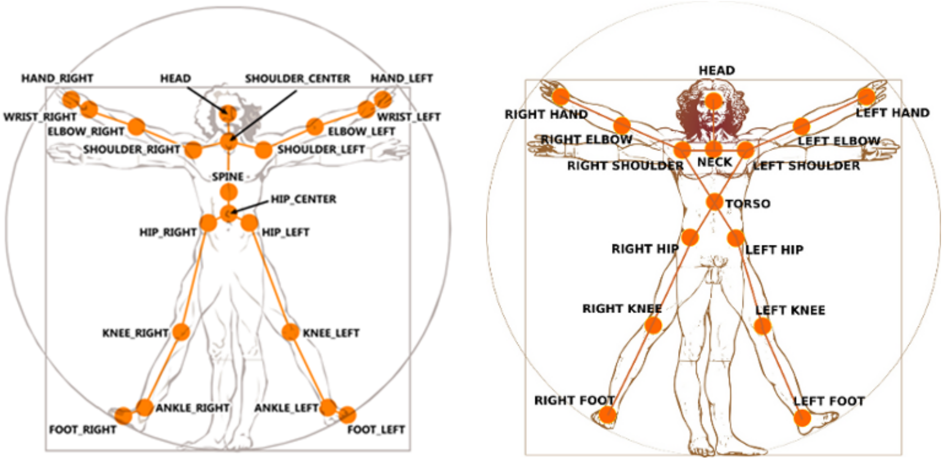


Figure 3.3: Location and name of the joints estimated by Microsoft (left) and OpenNI (right) skeletal trackers.

Another difference between the two skeletal trackers is about the ability in distinguishing almost frontal skeletons only (Microsoft) and also those seen from the back (OpenNI). With regards to the number of provided joints, the OpenNI's output has been adapted to the Microsoft's skeleton, adding five fictitious joints (wrists, ankles and hip center).

[1] *PrimeSense*'s patent: `http://patentscope.wipo.int/search/en/WO2007043036`

[2] `http://www.openni.org/`

Some example images from the dataset we used in the experiments recording with Microsoft SDK are reported in the first two rows in Figure 3.4, together with the estimated locations of twelve skeleton keypoints. At same figure, another set of example images is shown in the last two rows, in which the skeleton joints positions were estimated by OpenNI framework.

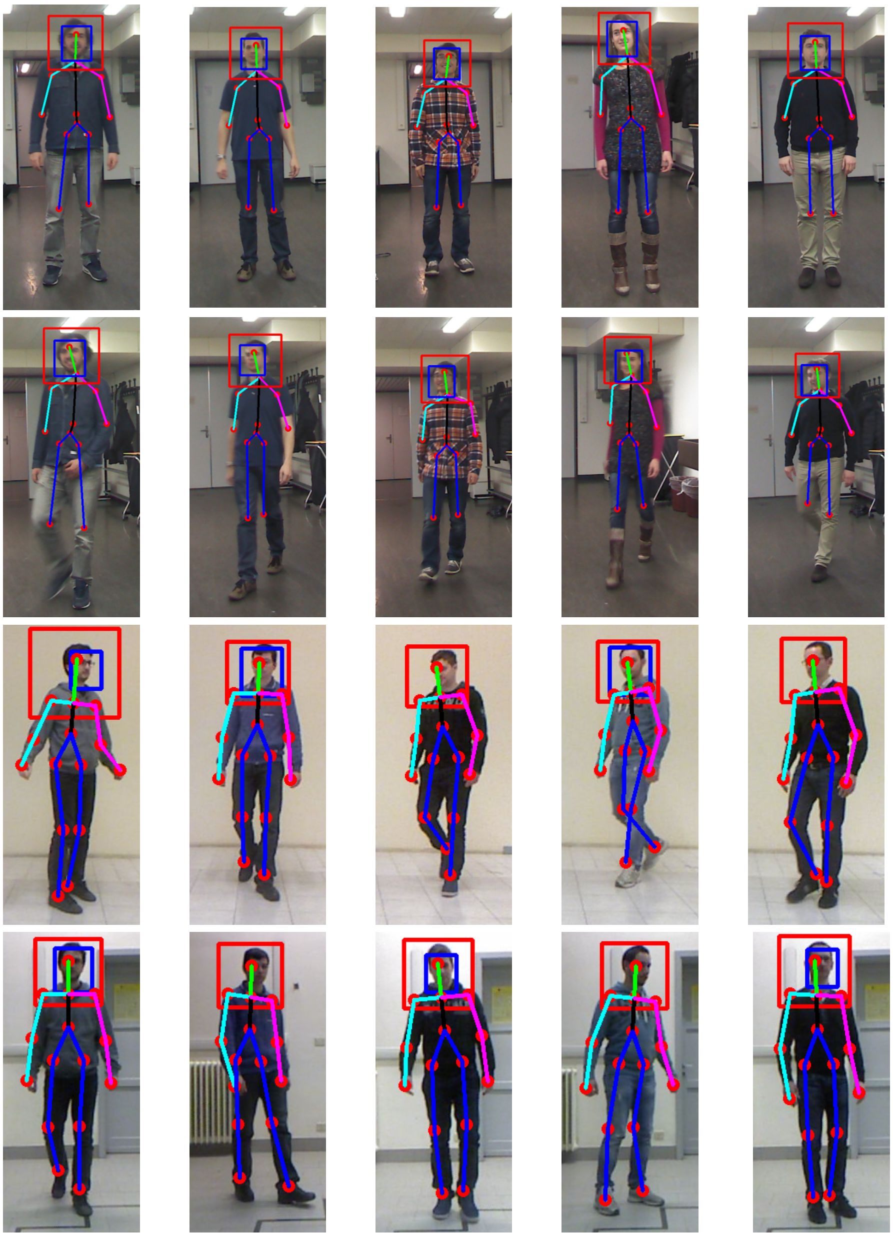


Figure 3.4: Example of training (first row) and testing (second row) images from the dataset in which Microsoft skeletal tracker was used. The last two rows show five example training (third row) and testing (fourth row) frames from the dataset created using OpenNI framework.

3.2 Skeleton-based Person Signature

For each frame, we extract a set of descriptors by using the human joints as keypoints. This choice comes from the following reasons: body joints are spread all over the human body, thus by extracting features from their locations we are sure to fully describe a person. Moreover, at every frame, keypoints are always concatenated in the same order, without the need for finding correspondences between keypoints belonging to training and testing images. In addition, the keypoint detection phase is not needed, thus saving a considerable amount of computational time.

In Figure 3.5, we show the *Skeleton-based Person Signature* (**SPS**) creation process for the first four skeleton keypoints. Local descriptors are drawn with a colored background and are concatenated into a single global signature which represents the person in the image.

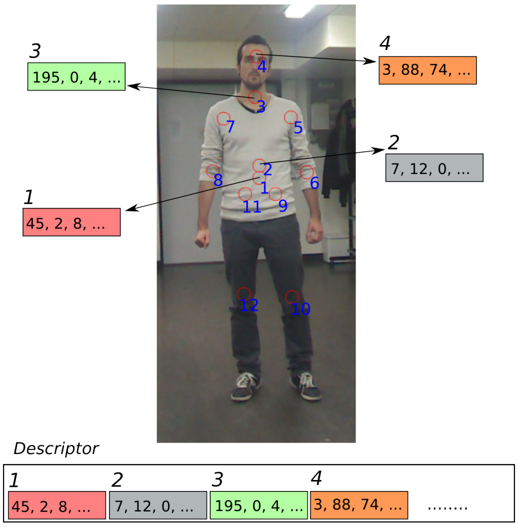


Figure 3.5: Example of Skeleton-based Person Signature creation: only the first four local descriptors are shown. In the final descriptor all local descriptors are juxtaposed. The red circles are centered on the keypoints and their radius are equal to the description region size.

In order to explain how the final **SPS** is generated, we need to first define the test τ which checks if a particular human joint is tracked or not:

$$\tau(\mathbf{J}_i; \mathbf{F}_k) := \begin{cases} 1 & \text{if } i\text{-th joint of frame } k \text{ is tracked} \\ 0 & \text{otherwise} \end{cases}, \tag{3.1}$$

where $i \in [0, M-1]$, $M = 20$ (total number of joints). Microsoft and OpenNI skeletal trackers automatically provide the value $\tau(\mathbf{J}_i; \mathbf{F}_k)$ for every joint.

In addition we define, only one time, a set of flags in which the values set to 0 are discarded even if $\tau(\mathbf{J}_i; \mathbf{F}_k) = 1$ for them and those set to 1 correspond to the joints involved in the **SPS** computation:

$$I(\mathbf{J}_i) := \begin{cases} 1 & \text{if } i\text{-th joint is involved in the } \textbf{SPS} \text{ computation} \\ 0 & \text{otherwise} \end{cases}, \tag{3.2}$$

In other words, setting $I(\mathbf{J}_i) = 1$, we assert that the **SPS** is computed concatenating also the descriptor vector extracted from the joint $\mathbf{J}_i$.

Now, let $D(\mathbf{J}_i; \mathbf{F}_k)$ be the local descriptor extracted around joint $\mathbf{J}_i$ at frame $\mathbf{F}_k$. Then, the final **SPS** for this frame is obtained by concatenating each $D(\mathbf{J}_i; \mathbf{F}_k)$. We tested our method by exploiting both 2D and 3D local descriptors in $D(\mathbf{J}_i; \mathbf{F}_k)$. All details about the **SPS** generation for the two mentioned cases follow in the next sections.

3.2.1 2D Signature

In our implementation, we set a description radius for $D(\mathbf{J}_i; \mathbf{F}_k)$ of 70 mm for both 2D and 3D features. For what concern 2D features, the radius in the image is variable with the distance from the sensor, thus the conversion from centimetres to pixels is obtained by exploiting depth information provided by the Kinect and the camera projection matrix, which allows to find the image point which is the projection of a 3D point of the scene. In symbols, given a point $\mathbf{T}(x, y, z)$, the projected point $\mathbf{P}(u, v)$ onto the image is obtained as:

$$\mathbf{p} = \mathbf{C}\,\mathbf{t} \ , \tag{3.3}$$

where $\mathbf{p}$ and $\mathbf{t}$ are the homogeneous representations of the points $\mathbf{P}$ and $\mathbf{T}$ respectively, and $\mathbf{C}$ is the camera projection matrix.

The **SPS** formula for the 2D case is slightly different depending on two situations which may occur: i) we considered only those frames in which $\tau = 1$ for each skeleton joint (Fully Visible Person, FVP); ii) we took account of the frame mentioned in i) and also of those in which only a subset of the joints $\mathbf{J}_i$ are tracked (Partially Occluded Person, POP).

3.2.1.1 Fully Visible Person

This configuration considers only the situation when the sensor records the entire set of body joints at frame k (all the joints are visible). Then, the **SPS** for these frames is calculated as follows:

$$\mathbf{SPS}_k = \left\{ \bigcup_{i=0}^{M-1} \Big(D(\mathbf{J}_i; \mathbf{F}_k) \mid I(\mathbf{J}_i) = 1 \Big) \right\} \Leftrightarrow \tau(\mathbf{J}_i; \mathbf{F}_k) = 1 \quad \forall i \in [0,\ M-1] \ , \tag{3.4}$$

where the union symbol stands for the juxtaposition of each local descriptor extracted around each joint considered for the **SPS** computation (as explained in equation 3.2).

Since OpenNI skeletal tracker allows to distinguish the frames in which the individuals are frontal from those in which they are seen from the back, a further flag stored this information when OpenNI framework was used. Instead, since Microsoft skeletal tracker is not able to correctly provide joints positions for a skeleton seen from the back, when it was exploited during the recordings, we discarded all the frames in which the individuals are walking in the opposite direction with respect to the Kinect sensor.

Three example frames for FVP configuration are shown in Figure 3.6 with their skeleton locations overlapped on the image. For this example configuration, $I(\mathbf{J}_i)$ was set to 1 for all the body joints.

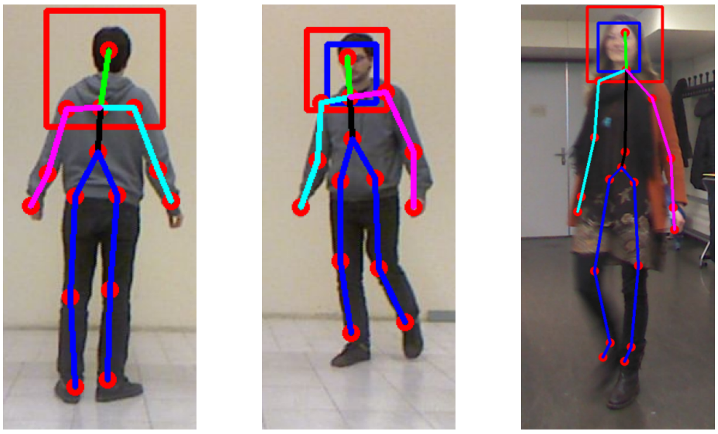


Figure 3.6: Three frames in which all the body joints are visible: back and frontal views for OpenNI framework (the first two figures) and for Microsoft one (the last figure).

3.2.1.2 Partially Occluded Person

In order to test the re-identification performance of our method in a more difficult situation, we decided to consider also the frames containing individuals with some of their body joints occluded: a typical example is when the persons show their profile, as in Figure 3.7, in which the right shoulder, for example, is occluded by the torso for both left and right example images.

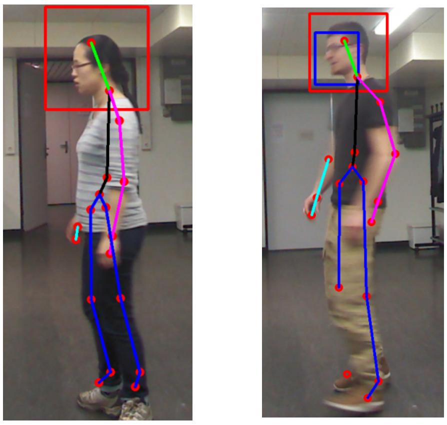


Figure 3.7: Two example frames of individuals walking in profile.

When POP configuration is considered, the formula for the **SPS** is different than the equation 3.4, because the check τ is now less restrictive:

$$\mathbf{SPS}_k = \bigcup_{i=0}^{M-1} \Big(D(\mathbf{J}_i; \mathbf{F}_k) \mid I(\mathbf{J}_i) \cdot \tau(\mathbf{J}_i; \mathbf{F}_k) = 1 \Big) \ . \tag{3.5}$$

In other words, this equation states that the **SPS** for a frame k is computed by juxtaposing only the local descriptors for those joints both visible ($\tau(\mathbf{J}_i;\mathbf{F}_k)=1$) and previously chosen for the **SPS** ($I(\mathbf{J}_i)=1$).

3.2.2 3D Signature

In addition to the 2D descriptors we also tested two different approaches based on 3D descriptors: i) the same configuration of the 2D case FVP (concatenation of local descriptors as explained in Section 3.2.1.1) but 3D descriptors with radius 70 mm are extracted from the person point cloud at the skeleton keypoints; ii) a global description of the person.

We did not implement a further method that considers also POP situation since the skeletal trackers we used are not good at providing 3D position estimates for the entire set of body joints when some of them are occluded. The **SPS** for the case i) is thus equal to the equation 3.4.

The approach ii) computes three 3D descriptors centered at the same keypoint but varies the description radius. In particular, the geometry of the points inside three spheres with radius 40, 80 and 120 cm is encoded in the descriptors, as shown in Figure 3.8.

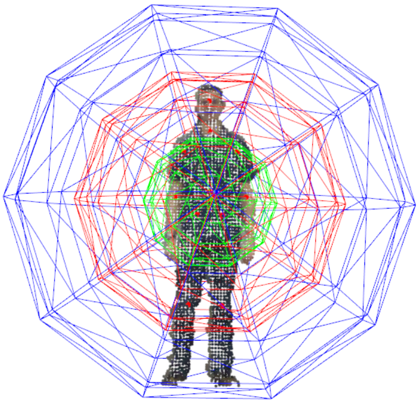


Figure 3.8: Three spheres with different radii centered at the hip center which represent the volume encoded by the global descriptors.

The three spheres are concentric and centered at the hip center. This method describes an individual by extracting descriptors from predetermined fixed areas. Thus, a person is characterized by the same portions of its body within all video sequences because the description radii are fixed. For this configuration the **SPS** is calculated as:

$$\mathbf{SPS}_k = \left\{ \bigcup_{r=0}^{R-1} \Big(D(\mathbf{J}_0;\mathbf{F}_k;r) \Big) \right\} \Leftrightarrow \tau(\mathbf{J}_i;\mathbf{F}_k)=1 \quad \forall i \in [0,\, M-1]\ , \tag{3.6}$$

where $R=3$ is the number of description radii, and $D(\mathbf{J}_0;\mathbf{F}_k;r)$ is the 3D descriptor extracted from the hip center ($\mathbf{J}_0$) with a description radius depending on the parameter r, as illustrated in Figure 3.8.

3.3 Matching strategy

Matching strategy looks for the correspondences between the individuals belonging to testing and training sets. The search is made by measuring the similarity among the Skeleton-based Person Signature (**SPS**) extracted from all frames of each individual in our datasets. Before explaining other details, we briefly introduce the theory behind the classification algorithm we used during our matching method.

3.3.1 K-Nearest Neighbor algorithm

K-nearest neighbor (KNN) is the simplest machine learning algorithm. It is a method to classify objects based on the majority class amongst its neighbors. Unlike most advanced classifiers, such as Support Vector Machine (SVM), KNN does not really generalize anything from training data and thus, no explicit training phase is needed. All the training elements are used during the testing step, when a decision is made based on the entire training set data.

Figure 3.9 shows an example of classification using a KNN algorithm (K is equal to 3 for this example). The classification problem considered in this figure, is to associate the circle object to the triangles or squares class. Since the majority of the objects among the K nearest ones belongs to the triangle class, the current testing object, the circle, is then classified as a triangle.

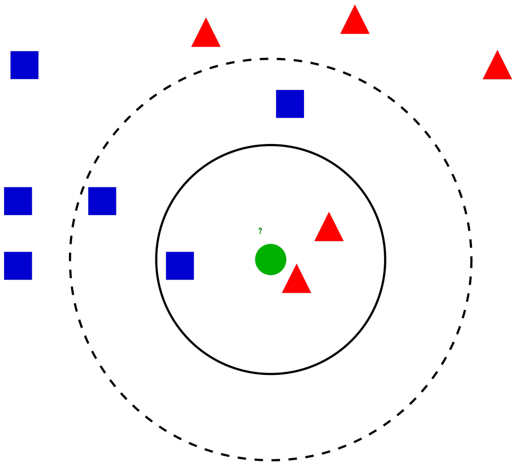


Figure 3.9: Knn example (K = 3). Image from `http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm`

As it can be seen in the above example, the choice of the parameter K can modify the classification results: in particular, if in Figure 3.9 we set K = 5, the classification output for the current testing object changes from triangle to square class. When K is set equal to 1, a special case for the KNN algorithm is performed: the classification is made only by looking for the nearest training class from the current testing object.

The computational cost of the KNN algorithm is linear with the training elements number, since it computes the distance of the testing object from each training element. When the number of training examples is very large, the time cost can be decreased to logarithmic if the search for finding the K nearest objects is computed on a KD-Tree (Section 3.3.2). Here, some portions of the searching space can be discarded based on the trees properties. First, the current nearest leaf is found and a scan nearby the leafs is performed; at some point, the distance from the query point to the leaf is higher than the worst point found so far. Thus, the nearest point is found and the search is stopped, because next leafs will not improve the search results.

3.3.2 KD-Tree

A KD-Tree is a data structure for efficiently storing and organizing a set of k-dimensional points. It is a particular binary tree and it is widely used in searching applications when the key is multidimensional. The use of KD-Trees improves the searching performance from $O(N)$ to $O(\log N)$, where N is the number of training classes, because some elements are not considered during the distance computation by exploiting some trees properties.

Each internal node of the tree generates an hyperplane perpendicular to a specific dimension that divides the space into two parts. Thus, at each level in the tree, some children are put to the left subtree and some others to the right subtree. The splitting is done at each level for a specific dimension. At the root of the tree, the division is based on the first dimension, at the next level the split is based on the second dimension and so on up to level K, where the division of the points is done based on the k-th dimension. At level K+1, the split is based at the first dimension again, and so on. A good choice to start the construction of the KD-Tree is the median point at the root, then based on the first dimension.

In Figure 3.10 there is an example about a 2D-Tree hyperplanes construction example for points (2,3), (5,4), (9,6), (4,7), (8,1), (7,2).

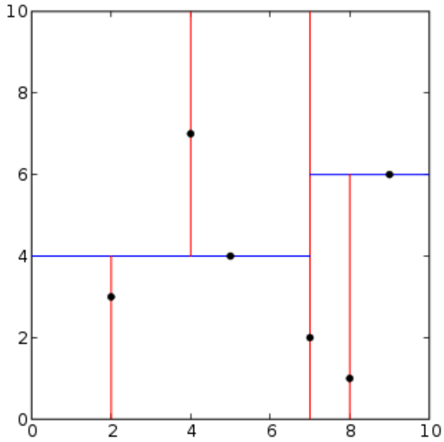


Figure 3.10: KD-Tree hyperplanes construction (K = 2). Image from `http://en.wikipedia.org/wiki/K-d_tree`

Figure 3.11 shows graphically the 2D-Tree generated starting from the hyperplanes constructed above in Figure 3.10.

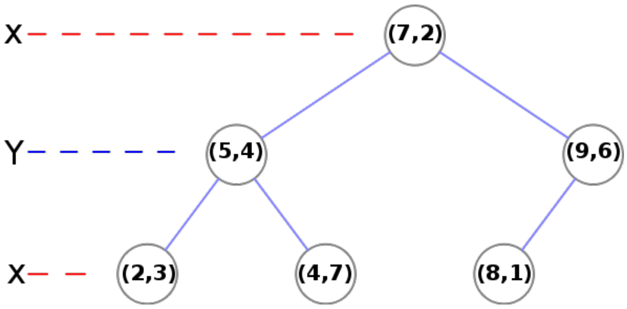


Figure 3.11: KD-Tree representation (K = 2). Image from `http://en.wikipedia.org/wiki/K-d_tree`

3.3.3 Skeleton-based Person Signature matching

At testing time, we compare the signature of a new testing frame with those extracted from the people training set by measuring descriptors similarity. Thus, we look for correspondences between individuals by performing a Nearest Neighbor search and using the Euclidean distance as similarity metric when signatures contain real values and the Hamming distance when signatures are composed by binary entries.

As mentioned in Section 3.2.1, the Person Signature extracted from a particular frame is different in case we only take account of frames in which all the body joints are seen from the sensor (FVP, Section 3.2.1.1) or in case we allow the partial description of the body (POP, Section 3.2.1.2). For the last case, at each new query signature, first we have to select from the training **SPS**s those for which at least the same N_k joints of the current frame k are available in the Person Signature, where N_k is computed as:

$$N_k = \sum_{i=0}^{M-1} \big(I(\mathbf{J}_i) \cdot \tau(\mathbf{J}_i; \mathbf{F}_k)\big)\,, \tag{3.7}$$

where $I(\mathbf{J}_i)$ and $\tau(\mathbf{J}_i; \mathbf{F}_k)$ were defined in Section 3.2.

Instead, when we consider the FVP configuration, no preliminary selection operations are needed. Anyway, the Euclidean distance between two Person Signatures is defined as:

$$d_E(\mathbf{F}_k; \mathbf{F}_w) = \sqrt{\sum_{m=0}^{L-1} \left[\mathbf{SPS}_k(m) - \mathbf{SPS}_w(m)\right]^2}\,, \tag{3.8}$$

where $\mathbf{SPS}_k(m)$ and $\mathbf{SPS}_w(m)$ are the m-th element of the skeleton-based signatures extracted from the testing frame $\mathbf{F}_k$ and the training frame $\mathbf{F}_w$, $L = Z \times Y$ is the total **SPS** length with Z equal to the feature descriptor size, i.e. $D(\mathbf{J}_i; \mathbf{F}_k)$ size, and $Y = \min(M, N_k)$ is the number of body keypoints involved in the **SPS**. In particular, $M = 20$ is the maximum number of available joints provided by the Microsoft skeletal tracker and N_k is computed in equation 3.7.

The Hamming distance [14] is defined as:

$$d_H(\mathbf{F}_k; \mathbf{F}_w) = \sum_{m=0}^{L-1} \left[\mathbf{SPS}_k(m) - \mathbf{SPS}_w(m)\right]^2, \qquad (3.9)$$

where the signatures $\mathbf{SPS}_k$ and $\mathbf{SPS}_w$ contain binary values only. In other words, the Hamming distance counts the number of positions where the corresponding bits are different.

Chapter 4

Experimental Results

For testing the re-identification methodology we explained in Chapter 3, we performed experiments on two publically available datasets: *BIWI RGBD-ID* and *IAS-Lab RGBD-ID*. The links to the datasets will be inserted soon on the *Intelligent Autonomous System Laboratory* (IAS-Lab) web page[1]. These datasets are targeted to long-term re-identification, thus people wear different clothes in training and testing videos. For our purposes, we selected from the datasets only the two testing videos acquired for every person, where people are wearing the same clothes. For what concerns the *BIWI RGBD-ID* dataset, we exploited the testing videos where the persons are still as our training set and the testing videos where the persons are walking as our testing set. Instead, the two testing videos of the *IAS-Lab RGBD-ID* dataset are recorded in two different rooms: we chose one of them as our training set and the other as our testing set. Moreover, the former dataset is composed by 28 persons and the latter by 11 individuals; at testing time, we assumed that all query people were present in the training set. Some example frames from the datasets are shown in Figure 3.4.

For evaluation purposes, we compute *Cumulative Matching Curves* (CMC) [28], which are commonly used for analyzing the re-identification performances. For every k from 1 to the number of training subjects, these curves express the mean person recognition rate computed when considering a classification to be correct if the ground truth person appears among the subjects who obtained the k best classification scores. The typical evaluation parameters for these curves are the *rank-1* recognition rate and the *normalized Area Under Curve* (nAUC), which is the integral of the CMC. In this work, the recognition rates are separately computed for every subject and then averaged to obtain the final recognition rate.

4.1 Testing configurations

We tested all the techniques explained in Section 3.2 using the two datasets mentioned above. Moreover, we also implemented and tested some state of the art techniques in order to compare the results achieved by our approach. Details about all these testing configurations follow in the next sections.

[1] `http://robotics.dei.unipd.it/`

4.1.1 Configurations for testing our approach

Concerning the *BIWI RGBD-ID* dataset, we implemented six different configurations which can be summarized as follows:

1. *2D-12k-1r-V*: 2D descriptors, 12 keypoints, 1 description radius, Fully Visible Person
2. *2D-12k-1r-O*: 2D descriptors, 12 keypoints, 1 description radius, Partially Occluded Person
3. *2D-20k-1r-V*: 2D descriptors, 20 keypoints, 1 description radius, Fully Visible Person
4. *2D-20k-1r-O*: 2D descriptors, 20 keypoints, 1 description radius, Partially Occluded Person
5. *3D-12k-1r-V*: 3D descriptors, 12 keypoints, 1 description radius, Fully Visible Person
6. *3D-1k-3r-V*: 3D descriptors, 1 keypoint, 3 description radii, Fully Visible Person

For the configuration with 12 keypoints we do not compute local descriptors for hands, wrists, ankles and foots, since usually they are very similar among all people. For this reason we decided to test the situation in which they are removed from the Person Signature, making it more discriminative and slightly faster to compute and match.

After evaluating the recognition performance on the *BIWI RGBD-ID* dataset, we decided to test our method also on the *IAS-Lab RGBD-ID* one, choosing from the list above only the configurations which reached the best re-identification score or that we considered to be a good trade-off between rank-1 score and computational cost. Since the tests showed that considering POP configuration (Section 3.2.1.2) the recognition performance gets worse, for the *IAS-Lab RGBD-ID* dataset we considered only FVP situation (3.2.1.1), testing a further case: i) only frontal people (F); ii) front-back people (FB):

1. *2D-12k-1r-F*: 2D descriptors, 12 keypoints, 1 description radius, Frontal People
2. *2D-12k-1r-FB*: 2D descriptors, 12 keypoints, 1 description radius, Front-Back People
3. *2D-20k-1r-F*: 2D descriptors, 20 keypoints, 1 description radius, Frontal People
4. *2D-20k-1r-FB*: 2D descriptors, 20 keypoints, 1 description radius, Front-Back People

4.1.2 Configurations for testing state of the art approaches

In order to compare the re-identification performances achieved by the approaches we developed with those exploited in other state of the art works, we implemented two further approaches: i) features extracted from keypoints provided by a detection algorithm; ii) color histograms.

About i), SIFT algorithm for both keypoint detection and description is used. In particular, for all frames, about sixty keypoints are detected, and then, described.

At testing time, a keypoint matching step is performed in order to find the correspondences between keypoints belonging to the testing and training frames. Thus, the similarity between two frames is given by the average Euclidean distance computed between the descriptors of each pair of keypoints.

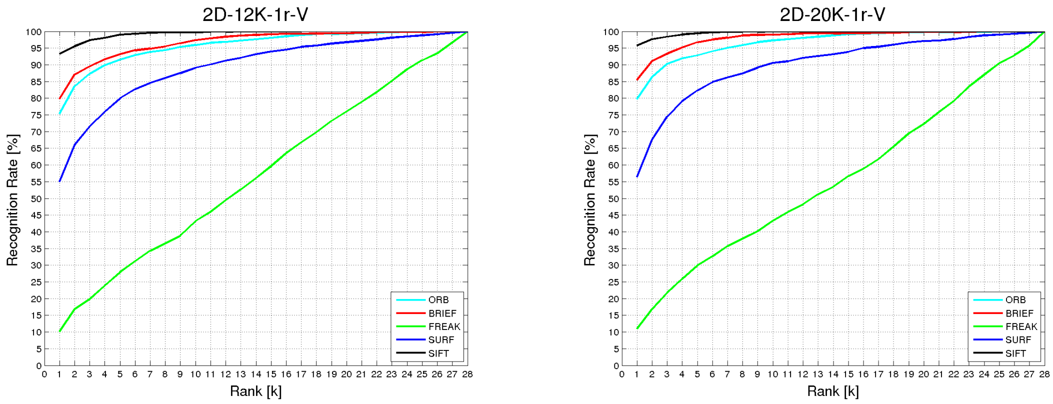
The method ii), relies on describing the people by encoding their body appearance exploiting color information. A color histogram for an image describes the color distribution in a specific color space, such as RGB or HSV. To compute the color histogram, the entire set of possible colors is divided into B ranges of values, called bins, and thus, the number of pixels that have colors in each bin is counted. RGB color space based histograms are not robust to illumination changes; for this purpose several works use a normalized color space in order to overcome this drawback. A widely used process is called *greyworld* normalization[29], in which each pixel of the current frame is normalized as: $red = R/(R + G + B)$, $green = G/(R + G + B)$, $blue = R/(R + G + B)$, R, G and B are the values of the color channels. In our experiments we assigned 32 bins for each color channel. Instead, when HSV color histograms are used, priority is generally given to the Hue component in order to ensure better illumination invariance. Indeed, in our tests we assigned 30 bins for Hue and 16 for Saturation and Value channels. Moreover, we also tested both Global and Local histograms: the former computes the color histogram of the entire appearance of the person while the latter computes one histogram for each body region. In particular we decided to describe torso, arms, forearms, upper and lower legs. The similarity between two histograms is based on the Bhattacharyya distance[13].

4.2 Tests on the BIWI RGBD-ID dataset

For this dataset the Microsoft SDK was exploited, recording RGB frames sized 1280 x 960 pixels and depth frames sized 640 x 480 pixels. The following sections illustrate the results for each configuration we tested.

4.2.1 Fully Visible Person configuration

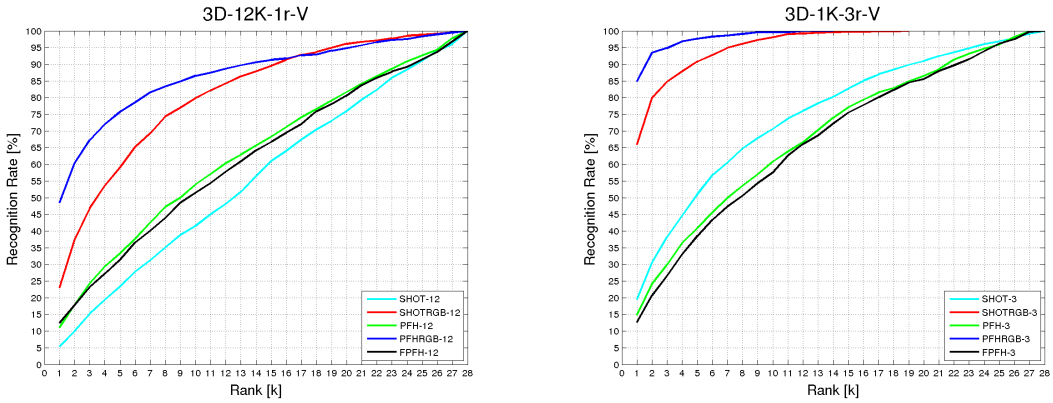
In these tests, descriptors computed on the image outperformed those computed on the point cloud. Figure 4.1b shows the configuration when all the 20 body joints are considered for the Signature computation. In Figure 4.1a the results for the situation when we limited to 12 the number of skeleton keypoints, thus discarding the descriptors for hands, wrists, ankles and foots, are illustrated. As it can be seen in Figure 4.1b, SIFT-20K-1r-V reached the best rank-1 and nAUC ($\text{rank}_{\text{sift}}$-1 = 95.75%, $\text{nAUC}_{\text{sift}}$ = 99.63%) for FVP configuration among all the descriptors we tested, but it resulted to be slower than other 2D descriptors: for example, it is 20 times slower than BRIEF (as illustrated in Table 4.1). This last descriptor is the best trade-off choice because it reaches a very high recognition performance (in Figure 4.1b, $\text{rank}_{\text{brief}}$-1 = 85.42%, $\text{nAUC}_{\text{brief}}$ = 98.15%) and its extraction time is only 10 ms. FREAK reaches the worst results for the 2D case ($\text{rank}_{\text{freak}}$-1 = 10.10%, $\text{nAUC}_{\text{freak}}$ = 57.54%): since its descriptor is based mostly on information at the center of the keypoint (saccadic search), it might not be robust to viewpoint and illumination changes.



(a) **(b)**

Figure 4.1: BIWI RGBD-ID dataset: CMCs for 2D descriptors when all the body joints are seen from the sensor. On the left: 12K configuration; on the right: 20K one.

For what concerns 3D descriptors, the 3D-1k-3r-V approach (Figure 4.2b) obtains better performance than the 3D-12k-1r-V one (Figure 4.2a). This means that the global shape and color are more significant than the local traits or that the noise introduced by the Kinect sensor is too big with respect to the local shape that should be encoded by local 3D descriptors. Moreover, the descriptors that capture also chromatic information outperform their corresponding version based on shape only. SHOT, PFH and FPFH are poor in terms of rank-1 score; PFHRGB, instead, achieves $\text{rank}_{\text{pfhrgb}}\text{-}1 = 84.84\%$ and $\text{nAUC}_{\text{pfhrgb}} = 98.65\%$, obtaining the best rank-1 score for the 3D case. However, for each frame 6500 ms are needed to extract this descriptor. SHOTRGB is the best combination between re-identification performance and execution time among 3D descriptors, with $\text{rank}_{\text{shotrgb}}\text{-}1 = 65.89\%$ and $\text{nAUC}_{\text{shotrgb}} = 95.88\%$: the time for extracting this descriptor from a frame is 65 ms only (100 times faster than PFHRGB).

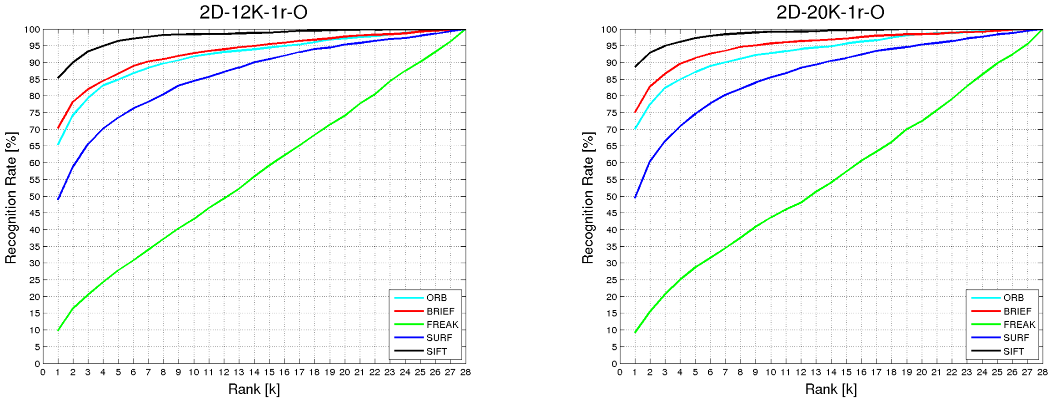


(a) **(b)**

Figure 4.2: BIWI RGBD-ID dataset: CMCs for 3D descriptors when all the body joints are seen from the sensor. On the left: 12K configuration; on the right: 1K and 3 description radii one.

4.2.2 Partially Occluded Person configuration

Figure 4.3 illustrates the CMCs for the POP configuration, in which the Person Signature is computed by considering both the FVP situation and also those frames containing some body joints not visible by the sensor, such as in Figure 3.7. In Figure 4.3a we considered the 12K configuration, while in Figure 4.3b all the body joints are involved in the Signature computation.



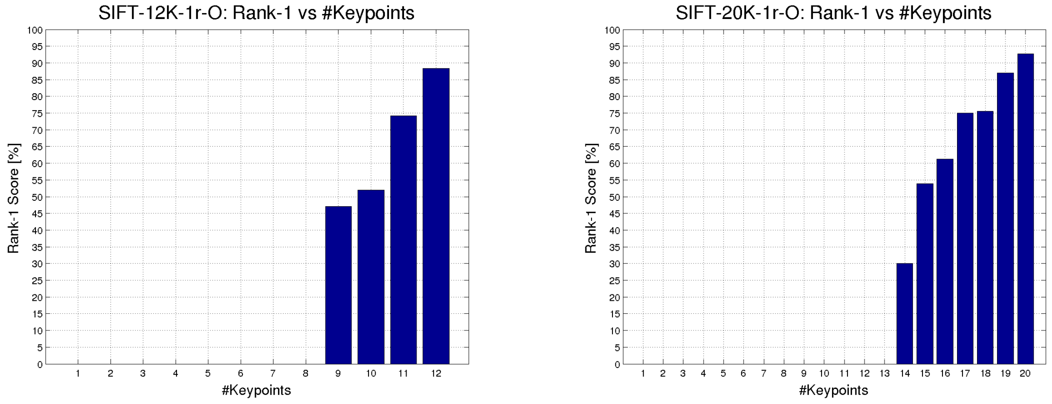
(a) (b)

Figure 4.3: BIWI RGBD-ID dataset: CMCs for 2D descriptors when not all the joints are seen from the sensor. On the left: 12K configuration; on the right: 20K one.

The CMCs above outline that the mean accuracy of the descriptors for POP configuration is slightly less with respect to the FVP configuration, but they remain in the same order for what concerns the rank-1 score: the configuration 20K (Figure 4.3b) achieves better performance in terms of rank-1 with respect to the 12K one (Figure 4.3a); SIFT reaches the best re-identification score and BRIEF is again a good trade-off choice.

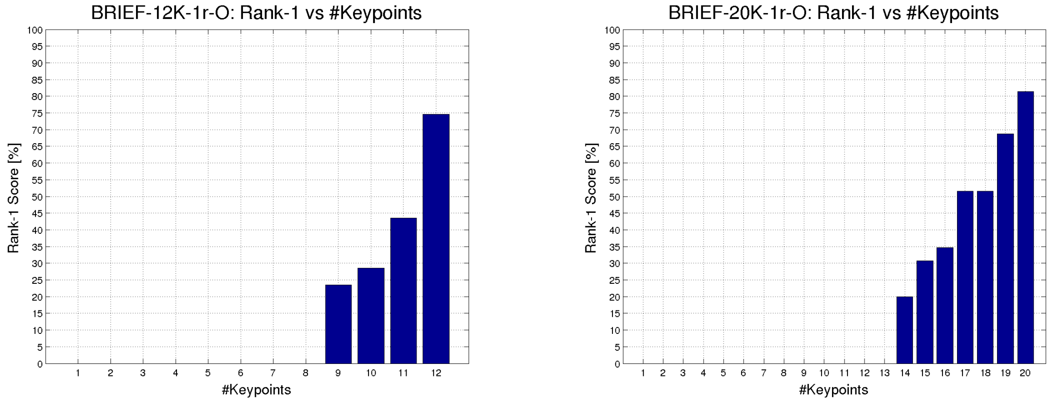
Figures 4.4 and 4.5 show, as histograms, the rank-1 score when varying the number of body joints involved in the **SPS** calculation. It can be seen that when the Person signature is extracted using a small number of body keypoints, such as when the individuals are seen from the side, the accuracy of the system is low. The Figure 4.4 is about SIFT based signature while Figure 4.5 is for BRIEF based signature.

In Figure 4.6 the percentage of frames generating the Person Signature for each number of body keypoints is shown. Here, it can be seen that the majority of the frames contains persons with all their body joints visible, but a small number of frames describes also individuals seen from the side (when the number of available keypoints decreases). The former situation achieves a higher accuracy than the latter and, for this reason, the FVP person configuration obtains better precision than the POP one, as illustrated in Figures 4.4 and 4.5.



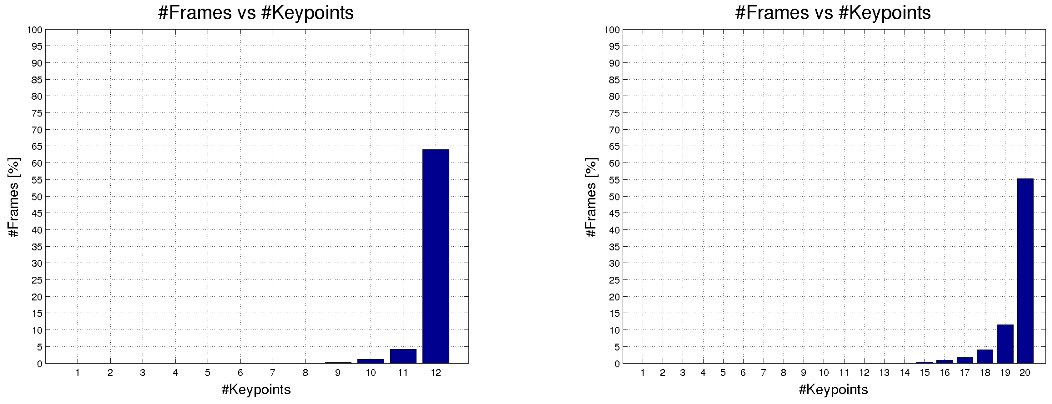
(a) (b)

Figure 4.4: Recognition score for SIFT based signature when varying the number of available keypoints. On the left, we limited to 12 the maximum number of keypoints (discarding hands, wrists, ankles and foots); on the right, no limitations regarding the keypoints.



(a) (b)

Figure 4.5: Recognition score for BRIEF bases signature when varying the number of available keypoints. On the left, we limited to 12 the maximum number of keypoints (discarding hands, wrists, ankles and foots); on the right, no limitations regarding the keypoints.

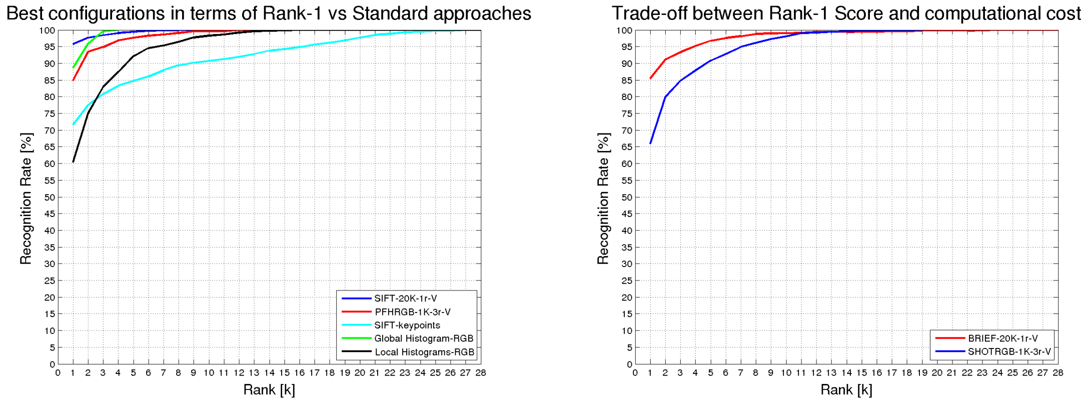


(a) (b)

Figure 4.6: Number of frames against number of keypoints used for the Person Signature. On the left: maximum 12 body keypoints (discarding hands, wrists, ankles and foots); on the right, no limitations regarding the keypoints.

4.2.3 Discussion

After evaluating the results obtained by each configuration we tested on the *BIWI RGBD-ID* dataset, we can derive some important considerations: 2D descriptors outperform the 3D ones; SIFT achieves the best performance in terms of rank-1 and BRIEF is to be considered the best trade-off choice between re-identification score and computational time (see also the Table 4.1). In order to compare our approaches with respect to other standard methods, we also implemented and tested two approaches widely used in literature, which do not exploit the skeleton keypoints: i) SIFT algorithm for both keypoint detection and description and ii) color histograms (Section 4.1.2). Our approach allows for better re-identification score than these last methods: to clarify graphically these considerations, in Figure 4.7a, the best CMCs in terms of rank-1 score obtained by both 2D and 3D descriptors (SIFT for the 2D case and PFHRGB for the 3D one) and the CMCs obtained by these further approaches are compared. Here, it can be seen that SIFT based Signature when considering all the 20 body joints and FVP configuration reaches the best classification score, while Global RGB histogram is the best state of the art approach for this dataset even if in terms of rank-1 score it is about 20 % lower with respect to the SIFT based signature. In Figure 4.7b, the best trade-off choices between recognition accuracy and computational load, choosing from all the 2D and 3D configurations, are drawn: BRIEF-20K-1r-V configuration for the 2D case and SHOTRGB-1K-3r-V configuration for the 3D one.



(a) Configurations and descriptors with highest rank-1

(b) Best trade-off between recognition accuracy and time cost

Figure 4.7: BIWI RGBD-ID dataset: on the left, configurations with highest rank-1; on the right, best trade-off choice between recognition accuracy and computational load.

Table 4.1 summarizes the rank-1 scores, together with the times needed for computing and matching descriptors when classifying one frame. Matching times have been estimated using a brute force algorithm, which finds the best match for a testing descriptor by evaluating its distance from each training descriptor. By exploiting KD-Trees and FLANN[2] based matcher during the Nearest Neighbor search, time performance improves about one order of magnitude. We do not report the tests

[2] `http://people.cs.ubc.ca/~mariusm/index.php/FLANN/FLANN`.

about the Partially Occluded configuration since it achieves lower accuracy than the Fully Visible one. Our tests were performed on an Intel$^{\circledR}$Core$^{\mathrm{TM}}$i3 CPU M330 @ 2.13 GHz with 4 GB DDR3 RAM.

Descriptor-configuration	*Extraction (ms)*	*Matching (ms)*	*Rank-1 (%)*
SIFT-12K-1r-V	**185.45**	**0.0045**	**93.27**
SURF-12K-1r-V	12.45	0.0072	55.01
BRIEF-12K-1r-V	**9.58**	**0.0024**	**79.82**
ORB-12K-1r-V	12.84	0.0024	75.35
FREAK-12K-1r-V	15.64	0.0048	10.10
SIFT-20K-1r-V	**309.08**	**0.0075**	**95.75**
SURF-20K-1r-V	20.75	0.012	56.46
BRIEF-20K-1r-V	**15.97**	**0.004**	**85.42**
ORB-20K-1r-V	21.4	0.004	79.74
FREAK-20K-1r-V	25.9	0.008	10.98
SHOT-12K-1r-V	616.09	0.0347	5.40
SHOTRGB-12K-1r-V	622.17	0.1482	23.01
PFH-12K-1r-V	765.37	0.0095	11.13
PFHRGB-12K-1r-V	**894.74**	**0.0137**	**48.49**
FPFH-12K-1r-V	1317.27	0.0021	12.43
SHOT-1K-3r-V	62.18	0.0045	19.49
SHOTRGB-1K-3r-V	**64.45**	**0.0179**	**65.89**
PFH-1K-3r-V	4086.20	0.0020	14.80
PFHRGB-1K-3r-V	**6527.03**	**0.0026**	**84.84**
FPFH-1K-3r-V	11196.59	0.0010	12.62
SIFT-keypoints	**413.28**	**0.0350**	**71.58**
Global RGB Histogram	**352.16**	**0.0075**	**88.59**
Local RGB Histogram	**309.15**	**0.0085**	**60.38**

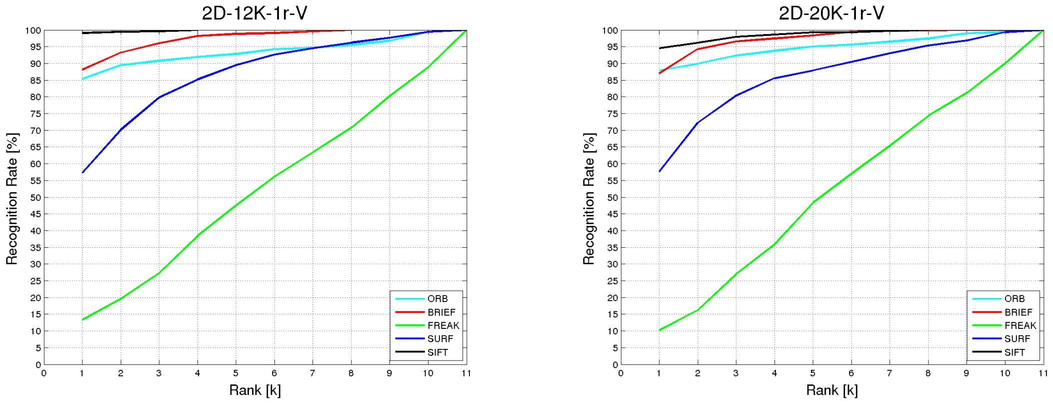
Table 4.1: Summary of rank-1 accuracy and computational times for descriptor extraction and matching: in orange, configurations with highest rank-1 score among 2D and 3D descriptors; in cyan, best trade-off between recognition performance and computational complexity; in purple, performance of the SIFT-based re-identification approach which does not exploit skeletal data (both keypoint detection and description are performed with the SIFT algorithm). In addition, in the last two rows, the results for Global and Local RGB Histograms are reported.

4.3 Tests on the IAS-Lab RGBD-ID dataset

In order to create this dataset we exploited the OpenNI framework, using the skeleton tracker algorithm provided by NiTE drivers. Differently from Microsoft's skeletal tracker, this one allows to distinguish frontal people from those seen from the back, giving us the opportunity for testing our method also in this situation. Thus, when we computed descriptor distances, we matched frontally obtained signatures one to each other and those obtained from the people seen from the back among them. Moreover, after evaluating the performances obtained by our approach on the *BIWI RGBD-ID* dataset, we decided to discard all frames in which not all the joints are visible, since the recognition accuracy does not improve with respect to the case in which all the body joints are available. We also decided to test only 2D descriptors because 3D ones are slower and less accurate. Another difference with respect to the *BIWI RGBD-ID* dataset is about the frame size: here, both RGB and depth images have 640 x 480 pixels.

4.3.1 Frontal Person configuration

For this tests, we only considered frontal people. In Figure 4.8a we reported the CMCs for 2D descriptors considering only 12 skeleton keypoints (discarding the descriptors for hands, wrists, ankles and foots), while in Figure 4.8b we included in the Person Signature each descriptor extracted from the entire set of the body joints.



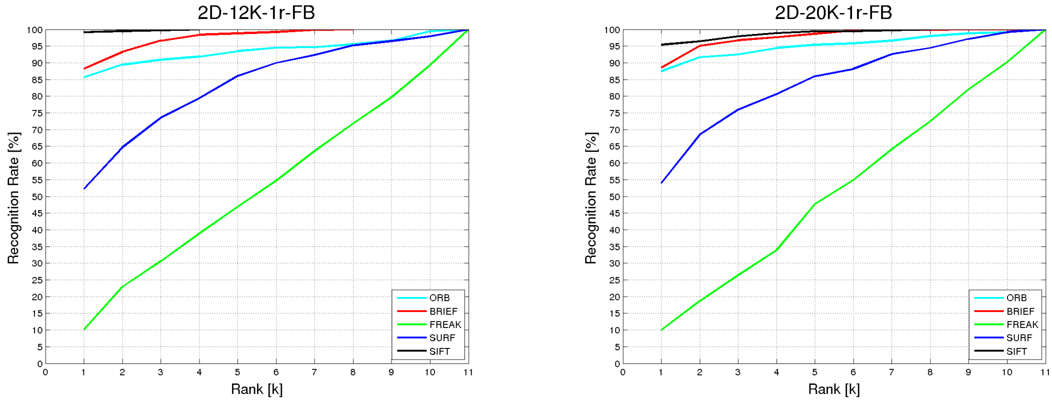
(a) (b)

Figure 4.8: IAS-Lab RGBD-ID dataset: CMCs for 2D descriptors when all the joints are seen frontally from the sensor. On the left: 12K configuration; on the right: 20K one.

By analyzing the results of the tests performed on the *IAS-Lab RGBD-ID* dataset, we can derive almost the same conclusions as for the *BIWI RGBD-ID* dataset. In particular, SIFT allows to reach the best classification score while BRIEF is the best trade-off choice between accuracy and computational cost, as also illustrated in Table 4.1. Furthermore, it can be noticed that, when extracting the Person Signature from 12 skeleton keypoints only, we obtained highest recognition rate than using all the 20 body joints. Since for this dataset we recorded frames sized 640 x 480 (a quarter than the frame size of *BIWI RGBD-ID* dataset) and the skeleton provided by OpenNI skeletal tracker is less stable than that given by Microsoft skeletal tracker, the noise introduced by including also the descriptors for hands, wrists, ankles and foot within the Person Signature are probably higher than the information that they add in the Signature.

4.3.2 Front-Back Person configuration

In a further case, we tested the re-identification accuracy when taking into account also those frames in which OpenNI skeleton tracker detects persons seen from the back: in Figure 4.9a, the CMCs when considering 12 skeleton keypoints are illustrated, while 20 keypoints case is shown in Figure 4.9b.



(a)

(b)

Figure 4.9: IAS-Lab RGBD-ID dataset: CMCs for 2D descriptors when all the joints are seen from the back by the sensor. On the left: 12K configuration; on the right: 20K one.

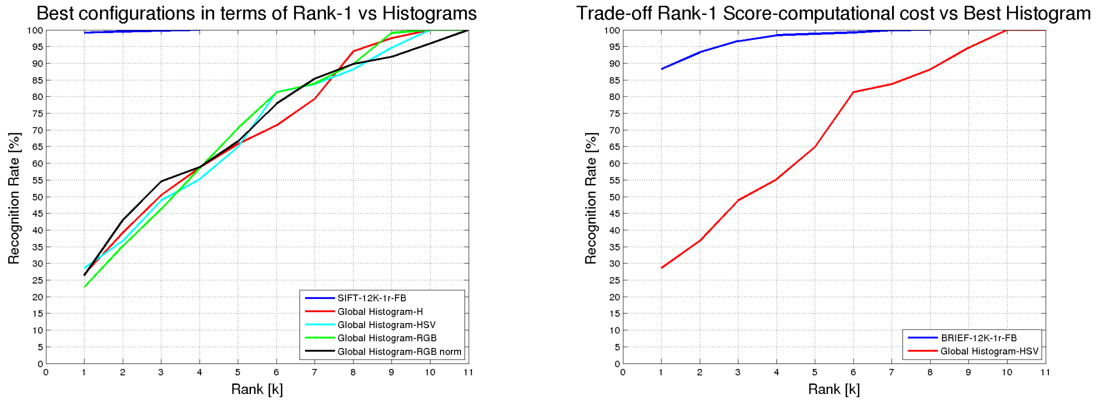
By allowing also Front-Back skeletons, the performances improve with respect to the Frontal people configuration, even if by comparing the above figures and those shown in Section 4.3.1 this fact is not easy to notice. In particular, SIFT-12K-1r-FB configuration achieves the best results (rank$_{\text{sift}}$-1 = 99.16%, nAUC$_{\text{sift}}$ = 99.85%) and, as for *BIWI RGBD-ID* dataset, the BRIEF based signature when 12K-1r-FB configuration is considered, represents the trade-off choice between accuracy and time load, reaching rank$_{\text{brief}}$-1 = 88.30%, nAUC$_{\text{brief}}$ = 97.68% and it is about 20 times faster than SIFT in extracting the descriptor for a frame (Table 4.1). Since the results obtained by Front-Back People configuration are better than the Frontal People one, illustrated in the section above, we can assert that our approach does not penalize the description of people seen from the back, thus we do not need for discarding these frames.

4.3.3 Discussion

Similarly to the results obtained from the *BIWI RGBD-ID* dataset, by analyzing the results illustrated for the *IAS-Lab RGBD-ID* dataset, we can assert that SIFT is the best choice when taking account of the rank-1 score only and BRIEF based signature is suitable if also the time cost for descriptor extraction is important (20 times faster). We do not report the table of the times needed for extracting and matching a frame by using all descriptors for each configuration, because they remain in the same ratio shown in table 4.1, but gaining in terms of absolute values an amount of time about 20 %. This gain is caused by the image resolution used for acquiring the *IAS-Lab RGBD-ID* dataset: it is a quarter with respect to the size of the images recorded in the *BIWI RGBD-ID* dataset and thus, at the feature extraction step, a lower number of pixels are involved during the descriptor vector computation, even if the the description radius is the same.

In Figure 4.10a, the CMC for the best configuration in terms of re-identification score is plotted, that is the SIFT based descriptor limited to 12 human joints in the Person Signature. Here, also the curves for some further approaches based on global color histograms, mentioned in Section 4.1.2, are reported in order to compare the result achieved by our method with respect to standard state of the art techniques which widely use color based signatures. It can be seen that our approach reaches better re-identification accuracy.

Figure 4.10b illustrates the best trade-off between accuracy and time cost. Thus, in this figure we compared BRIEF based signature with respect to the best histogram based approach (HSV global histogram).



(a) SIFT vs Global Histograms

(b) BRIEF vs HSV histogram

Figure 4.10: IAS-Lab RGBD-ID dataset: on the left: best descriptors in terms of rank-1; on the right: best trade-off choice.

A standard state of the art method which computes RGB global histograms obtained the worst result in this context, reaching rank_{RGB}-1 = 22.99% and nAUC_{RGB} = 71.60%. For this reason, differently from *BIWI RGBD-ID* dataset, we implemented and tested three further histogram based approaches: RGB-norm, HSV and Hue only, explained in Section 4.1.2. We observed that none of these approaches reaches good re-identification performance, as shown in Figure 4.10a, remarking another time the excellent results we achieved by our approach which exploits skeleton keypoints.

4.4 Comparison between the BIWI RGB-ID and the IAS-Lab RGBD-ID datasets results

As mentioned in the above sections, the differences between these datasets are about the skeleton tracking algorithm used, the number of individuals which are present within the training and testing sets and the size of the RGB images. An important fact we observed is about the stability of the skeletal trackers: when we exploited Microsoft skeletal tracking algorithm, the body joints are almost detected in the same body locations; instead, for what concerns the OpenNI framework, we can state that it is more noisy. This fact is also confirmed by analyzing the results of the 12K and 20K configurations we tested on both *BIWI RGB-ID* and *IAS-Lab RGBD-ID* datasets: the former configuration discards the descriptors of body extrema (hands, wrists, ankles and foots) while the latter considers all the body joints for computing the Person Signature. In an ideal situation, each joint adds distinctive information; indeed, for the *BIWI RGB-ID* dataset, the 20K configuration overcome the 12K one, while with regard to the *IAS-Lab RGBD-ID* dataset, we experienced the contrary. For this reason, we can conclude that Microsoft skeletal tracker is more accurate than the one provided by OpenNI framework. The number of recorded people affects the recognition score because in the *BIWI RGB-ID* dataset, the nearest people are to be searched among 28 individuals while, in the *IAS-Lab RGBD-ID* dataset, among 11.

About the accuracy of the descriptors we tested, the results obtained by our approach on the *BIWI RGB-ID* dataset confirm that 2D descriptors allow to reach higher classification score and they are faster than the 3D ones. Among 2D descriptors, SIFT based signature achieves the best rank-1 score but it is slower than the other 2D descriptors; thus, BRIEF based signature is the best trade-off choice between accuracy and time cost.

Moreover, the comparison between the results obtained by our method and standard approaches, such as those based on SIFT keypoints or color histograms (Section 4.1.2), highlights the excellent re-identification score we obtained describing people by exploiting the body joints as keypoints.

4.5 Searching other RGB-D datasets for people re-identification

Convinced by the good results we obtained, we then decided to prove our approach on other publically available RGB-D datasets for people re-identification. Depth information is often needed for obtaining a real time estimation of the skeleton of a person, which is a fundamental information for our approach. With regard to long-term people re-identification, the *RGB-D Person Re-identification*[3] dataset is a good choice, because it is composed by 79 individuals and some of them change clothes between training and testing videos. Instead, we noticed that currently it does not exist any dataset suitable for our purpose, that is targeted to short-term

[3] `http://www.iit.it/en/datasets/rgbdid.html`

people re-identification with RGB-D data in which the individuals wear the same clothes among training and testing frames. In literature, a large number of datasets, composed by RGB frames only, was exploited for the people re-identification task: the most widely used is the *VIPeR* dataset[30]. When depth information is not available, histogram-based approaches are usually used, by assuming that a person has almost the same appearance from each viewpoint. Our approach, instead, requires to know the positions of the skeleton keypoints; moreover, to the best of our knowledge, it does not exist any skeletal tracker able to estimate in real time the body joints by exploiting RGB information only, thus without using the depth values. We were therefore unable to test our method by using other datasets than those mentioned in Section 4.

Chapter 5

Conclusions

In this work, a novel method for people re-identification which extracts feature descriptors from the joints locations of the human body is presented. The standard state of the art approaches rely on describing the individuals by exploiting keypoints provided by a detection algorithm or by comparing color histograms of the body appearance. Instead, in this work, Person Signatures generated by concatenating in a specific order each body skeleton descriptor are computed and matched. Thus, keypoint matching process is unnecessary because the correspondences between pairs of keypoints belonging to training and testing frames are already known.

We proved that this approach allows to increase the recognition rate and decrease the computational time with respect to standard techniques. Moreover, we tested and compared a number of state of the art 2D and 3D feature descriptors on two publically available datasets and with several different approaches. We also tested two skeletal trackers for providing the positions of the body joints: Microsoft's and OpenNI's. In our tests, the SIFT based Person Signature resulted to reach the best re-identification score while the BRIEF based Person Signature guarantees the best trade-off between accuracy and computational cost.

As future works, we will test our method when exploiting other types of feature classifiers (e.g. SVM) for the matching phase. Furthermore, our long-term purpose is to perform the re-identification as an on-line application integrated in a tracking framework.

Bibliography

[1] Kristin Koch, Judith McLean, Ronen Segev, Michael A Freed, Michael J Berry II, Vijay Balasubramanian, and Peter Sterling. How much the eye tells the brain. *Current Biology*, 16(14):1428–1434, 2006.

[2] Kyongil Yoon, David Harwood, and Larry Davis. Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation (JVCIR 2006)*, 17(3):605–622.

[3] Dung-Nghi Truong Cong, C. Achard, and L. Khoudour. People re-identification by classification of silhouettes based on sparse representation. In *Proceedings of the 2nd International Conference on Image Processing Theory Tools and Applications (IPTA 2010)*, pages 60–65.

[4] Lei Hu, Shuqiang Jiang, Qingming Huang, and Wen Gao. People re-detection using adaboost with sift and color correlogram. In *Proceedings of the 15th International Conference on Image Processing (ICIP 2008)*, pages 1348–1351.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2360–2367.

[6] Kai Jungling and Michael Arens. Feature based person detection beyond the visible spectrum. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2009)*, pages 30–37.

[7] I.O. de Oliveira and J.L. de Souza Pio. People reidentification in a camera network. In *Proceedings of the 8th IEEE Conference on Dependable, Autonomic and Secure Computing (DASC 2009)*, pages 461–466.

[8] D. Baltieri, R. Vezzani, R. Cucchiara, A. Utasi, C. Benedek, and T. Sziranyi. Multi-view people surveillance using 3d information. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops 2011)*, pages 1817–1824.

[9] Wu Liu, Tian Xia, Ji Wan, Yongdong Zhang, and Jintao Li. Rgb-d based multi-attribute people search in intelligent visual surveillance. In *Advances in Multimedia Modeling (AMM 2012)*, pages 750–760. Springer.

[10] J. Oliver, A. Albiol, and A. Albiol. 3d descriptor for people re-identification. In *Proceedings of the 21st IEEE International Conference on Pattern Recognition (ICPR 2012)*, pages 1395–1398.

[11] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *Proceedings of the 2012 European Conference on Computer Vision Workshops and Demonstrations (ECCV Workshops and Demonstrations 2012)*, pages 433–442. Springer.

[12] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics 1951*, 22(1):79–86.

[13] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc. 1943*, 35(1):99–109.

[14] Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal 1950*, 29(2):147–160.

[15] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV 1999)*, volume 2, pages 1150–1157.

[16] Yan Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages II–506–II–513 Vol.2.

[17] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

[18] Luo Juan and Oubong Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP 2009)*, 3(4):143–152.

[19] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June.

[20] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: binary robust independent elementary features. In *Proceedings of the 2010 European Conference on Computer Vision (ECCV 2010)*, pages 778–792. Springer.

[21] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV 2011)*, pages 2564–2571.

[22] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pages 510–517.

[23] R.B. Rusu, N. Blodow, Z.C. Marton, and M. Beetz. Aligning point cloud views using persistent feature histograms. In *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008)*, pages 3384–3391.

[24] R.B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA 2009)*, pages 3212–3217.

[25] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Proceedings of the 2010 European Conference on Computer Vision (ECCV 2010)*, pages 356–369. Springer.

[26] Federico Tombari, Samuele Salti, and Luigi Di Stefano. A combined texture-shape descriptor for enhanced 3d feature matching. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP 2011)*, pages 809–812. IEEE.

[27] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 1297–1304.

[28] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the 2008 European Conference on Computer Vision (ECCV 2008)*, volume 5302 of *Lecture Notes in Computer Science*, pages 262–275. Springer Berlin Heidelberg.

[29] José M Buenaposada and Luis Baumela. Variations of grey world for face tracking. *Image Processing and Communications*, 7(3-4):51–62, 2001.

[30] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the 2008 European Conference on Computer Vision (ECCV 2008)*, pages 262–275. Springer.