
**Universitat Politècnica de València**

# P H D  T H E S I S

# Person Re-Identification using RGB-Depth Cameras

Defended by

Javier Oliver Moll

Thesis Advisor: Alberto Albiol Colomer

November 2015

## Acknowledgments

To my family.

**Person Re-Identification using RGB-Depth Cameras**

**Abstract:**

The presence of surveillance systems in our lives has drastically increased during the last years. Camera networks can be seen in almost every crowded public and private place, which generate huge amount of data with valuable information. The automatic analysis of data plays an important role to extract relevant information from the scene. In particular, the problem of person re-identification is a prominent topic that has become of great interest, specially for the fields of security or marketing. However, there are some factors, such as changes in the illumination conditions, variations in the person pose, occlusions or the presence of outliers that make this topic really challenging. Fortunately, the recent introduction of new technologies such as depth cameras opens new paradigms in the image processing field and brings new possibilities. This Thesis proposes a new complete framework to tackle the problem of person re-identification using commercial rgb-depth cameras. This work includes the analysis and evaluation of new approaches for the modules of segmentation, tracking, description and matching. To evaluate our contributions, a public dataset for person re-identification using rgb-depth cameras has been created.

Rgb-depth cameras provide accurate 3D point clouds with color information. Based on the analysis of the depth information, an novel algorithm for person segmentation is proposed and evaluated. This method accurately segments any person in the scene, and naturally copes with occlusions and connected people. The segmentation mask of a person generates a 3D person cloud, which can be easily tracked over time based on proximity.

The accumulation of all the person point clouds over time generates a set of high dimensional color features, named raw features, that provides useful information about the person appearance. In this Thesis, we propose a family of methods to extract relevant information from the raw features in different ways. The first approach compacts the raw features into a single color vector, named Bodyprint, that provides a good generalisation of the person appearance over time. Second, we introduce the concept of 3D Bodyprint, which is an extension of the Bodyprint descriptor that includes the angular distribution of the color features. Third, we characterise the person appearance as a bag of color features that are independently generated over time. This descriptor receives the name of Bag of Appearances because its similarity with the concept of Bag of Words. Finally, we use different probabilistic latent variable models to reduce the feature vectors from a statistical perspective. The evaluation of the methods demonstrates that our proposals outperform the state of the art.

## Re-identificación de personas usando cámaras RGB-profundidad

**Resumen:**

La presencia de sistemas de vigilancia se ha incrementado notablemente en los últimos años. Las redes de videovigilancia pueden verse en casi cualquier espacio público y privado concurrido, lo cual genera una gran cantidad de datos de gran valor. El análisis automático de la información juega un papel importante a la hora de extraer información relevante de la escena. En concreto, la re-identificación de personas es un campo que ha alcanzado gran interés durante los últimos años, especialmente en seguridad y marketing. Sin embargo, existen ciertos factores, como variaciones en las condiciones de iluminación, variaciones en la pose de la persona, oclusiones o la presencia de artefactos que hacen de este campo un reto. Afortunadamente, la introducción de nuevas tecnologías como las cámaras de profundidad plantea nuevos paradigmas en la visión artificial y abre nuevas posibilidades. En esta Tesis se propone un marco completo para abordar el problema de re-identificación utilizando cámaras rgb-profundidad. Este trabajo incluye el análisis y evaluación de nuevos métodos de segmentación, seguimiento, descripción y emparejado de personas. Con el fin de evaluar las contribuciones, se ha creado una base de datos pública para re-identificación de personas usando estas cámaras.

Las cámaras rgb-profundidad proporcionan nubes de puntos 3D con información de color. A partir de la información de profundidad, se propone y evalúa un nuevo algoritmo de segmentación de personas. Este método segmenta de forma precisa cualquier persona en la escena y resuelve de forma natural problemas de oclusiones y personas conectadas. La máscara de segmentación de una persona genera una nube de puntos 3D que puede ser fácilmente seguida a lo largo del tiempo.

La acumulación de todas las nubes de puntos de una persona a lo largo del tiempo genera un conjunto de características de color de grandes dimensiones, denominadas características base, que proporcionan información útil de la apariencia de la persona. En esta Tesis se propone una familia de métodos para extraer información relevante de las características base. La primera propuesta compacta las características base en un vector único de color, denominado Bodyprint, que proporciona una buena generalización de la apariencia de la persona a lo largo del tiempo. En segundo lugar, se introducen los Bodyprints 3D, definidos como una extensión de los Bodyprints que incluyen información angular de las características de color. En tercer lugar, la apariencia de la persona se caracteriza mediante grupos de características de color que se generan independientemente a lo largo del tiempo. Este descriptor recibe el nombre de Grupos de Apariencias debido a su similitud con el concepto de Grupos de Palabras. Finalmente, se proponen diferentes modelos probabilísticos de variables latentes para reducir los vectores de características desde un punto de vista estadístico. La evaluación de los métodos demuestra que nuestras propuestas superan los métodos del estado del arte.

**Re-identificació de persones amb càmeres RGB-profunditat**

**Resum:**

La presència de sistemes de vigilància s'ha incrementat notòriament en els últims anys. Les xarxes de videovigilància poden veure's en quasi qualsevol espai públic i privat concorregut, la qual cosa genera una gran quantitat de dades de gran valor. L'anàlisi automàtic de la informació pren un paper important a l'hora d'extraure informació rellevant de l'escena. En particular, la re-identificaciò de persones és un camp que ha aconseguit gran interès durant els últims anys, especialment en seguretat i màrqueting. No obstant, hi ha certs factors, com variacions en les condicions d'il.luminació, variacions en la postura de la persona, oclusions o la presència d'artefactes que fan d'aquest camp un repte. Afortunadament, la introducció de noves tecnologies com les càmeres de profunditat, planteja nous paradigmes en la visió artificial i obri noves possibilitats. En aquesta Tesi es proposa un marc complet per abordar el problema de la re-identificació mitjançant càmeres rgb-profunditat. Aquest treball inclou l'anàlisi i avaluació de nous mètodes de segmentació, seguiment, descripció i emparellat de persones. Per tal d'avaluar les contribucions, s'ha creat una base de dades pública per re-identificació de persones emprant aquestes càmeres.

Les càmeres rgb-profunditat proporcionen núvols de punts 3D amb informació de color. A partir de la informació de profunditat, es defineix i s'avalua un nou algorisme de segmentació de persones. Aquest mètode segmenta de forma precisa qualsevol persona en l'escena i resol de forma natural problemes d'oclusions i persones connectades. La màscara de segmentació d'una persona genera un núvol de punts 3D que pot ser fàcilment seguida al llarg del temps.

L'acumulació de tots els núvols de punts d'una persona al llarg del temps genera un conjunt de característiques de color de grans dimensions, anomenades característiques base, que hi proporcionen informació útil de l'aparença de la persona. En aquesta Tesi es proposen una família de mètodes per extraure informació rellevant de les característiques base. La primera proposta compacta les característiques base en un vector únic de color, anomenat Bodyprint, que proporciona una bona generalització de l'aparença de la persona al llarg del temps. En segon lloc, s'introdueixen els Bodyprints 3D, definits com una extensió dels Bodyprints que inclouen informació angular de les característiques de color. En tercer lloc, l'aparença de la persona es caracteritza amb grups de característiques de color que es generen independentment a llarg del temps. Aquest descriptor reb el nom de Grups d'Aparences a causa de la seua similitud amb el concepte de Grups de Paraules. Finalment, es proposen diferents models probabilístics de variables latents per reduir els vectors de característiques des d'un punt de vista estadístic. L'avaluació dels mètodes demostra que les propostes presentades superen als mètodes de l'estat de l'art.

# Contents

Chapter 1

# Introduction

## Contents



The use of camera networks has widespread during recent years. Every day, billions of Closed-Circuit TeleVision cameras (CCTV) are being used for surveillance applications. A recent study reveals that only in United Kingdom there are aproximately 2 million CCTV cameras [Gerrard n 42], where more than 100.000 are concentrated in London.

Vision-based systems have become essential in domains such as surveillance, marketing, gaming or sports. Systems may be located in public spaces, such as shopping areas or crowded town centres, in transport infrastructures such as rail stations and airports, or in private areas such as sports/conference venues or retail stores, just to enumerate a few. The ubiquity of these vision systems allow to monitor people activity and obtain information about people behavior, occupacy, trajectory, and also check security areas or abandoned objects.

However, several problems arise with the widespread of the vision-based systems. It is obvious that manually processing such a huge amount of data to survey the camera network turns to be unfeasible for dedicated staff members in control rooms (see Figure 1.1). According to neuroscience models, people can only focus their attention in particular regions of their field of vision for a specific instant. Green [Green 1999] claims that operators significantly drop their concentration after 20 minutes of image analysis. Therefore, only a small portion of the cameras can be correctly analysed in real-time. In addition to human limitations, vision-based surveillance has been widely criticized on several grounds such as violations of the privacy of the people, illegality and preventing social freedoms.

These facts strongly motivate the need of automatic video analysis that would provide higher level information to the human experts only when the situation turned critical. Operators can take more rapid and more effective decisions when

Figure 1.1: An example of CCTV control room at United College [operators ].

analysing such high-level information cues. This fact justifies the clear trend in vision-based surveillance research to automate this process. For example, the U.S. government has recently funded the development of an automatic video surveillance technology [Oltramari 2012] to detect and report illicit behavior all around the country. In order to be able perform such an automatic analysis of the scene, the task of *person re-identification* becomes the key aspect for all vision-based systems and can be applied to all domains.

## 1.1 Problem description

The task of person re-identification is the task of recognizing an individual from a large set of candidates, who can be seen in the same or different place at a different time. This task has gained relevance in many different domains such as surveillance, but in particular in shopping malls and shops not only for security but also for marketing purposes.

Marketing managers for shopping malls rely on visitor statistics to measure the impact of their marketing campaign. The scientific measurement of marketing effectiveness is achieved by combining sales information with visitor statistics, which is superior to traditional methods that only take into account sales data. Some of the marketing metrics are CPM (Cost Per Thousand) and SSF (Shoppers per Square Foot), and identify the rent according to the total number of visitors to the mall or according to the number of visitors to each individual store in the mall.

People tracking in shops is a key point since managers can identify hot spots

and see what parts of the shop arouse more interest among the clients, so that the manager can adopt measures to benefit from this information. Controlling the occupacy is also an important task to ensure that the building is below the safe level of occupacy so as to avoid any massification. Besides, people counting in these retail environment is necessary if managers want to calculate the above mentioned conversion rates. Managers may also find interesting to find how long a person stays in the shop to evaluate the effect of certain campaign.

The question that arises here is: how can a vision-based system perform people re-identification? The problem of re-identification for a human being turns to be a simple and innate process that human can easily perform without previously thinking how to do it. The human visual system can effectively process the environment and just focus its attention to the relevant information from the scene. This skill of humans arises from years of implicit training, where people have learnt the nature of things. However, this process is not straightforward for a non-human. From the very beginings of the computer vision, researchers have been trying to replicate the way the human visual system works in order to be able to effectively describe the world just as a human brain does [Marr 67 8] [Gibson 9598]. But the point is: are the inputs the same for the human visual system and for computer vision machines?

The process of person re-identification in computer vision is generally tackled by taking one or several images of a person at a particular camera and extracting their global appearance, which will be compared with other descriptors belonging to the people detected by the same or different cameras at different time. Up to now, the process of people re-identification has only been addressed from a 2D point of view. Although color images can almost fully describe a scene, there is still some important information that is missing, as the 3D information.

The introduction of rgb-depth cameras has brought new opportunities (and also new challenges) to researchers in image processing. With these cameras researchers can go beyond the standard techniques to provide more accurate high-level descriptions of the scene. This new technology is changing day by day the concept of classic image processing.

In this Thesis, different novel methods for person re-identification in real and complex scenarios are introduced and discussed. In particular, the Thesis addresses new robust algorithms for people segmentation, description and new matching strategies making use of rgb-depth sensors in a non-overlapping camera system.

## 1.2 Challenges of People Re-Identification

Multiple person re-identification is an open-set matching problem with a dynamically-evolving and unconstrained gallery set. Matching people in real scenarios under no constrains is a challenging task since both people appearance and environment can strikingly change over time and also among cameras.

Challenges in people re-identification using appearance information can be classified into two groups, depending on the nature of the causing factors. On the

one hand, the concept of intrinsic factors can be introduced to describe all those variations directly generated by a person. On the other hand, the extrinsic factors describe all the environmental phenomenons that can change the perception of the person appearance captured by a camera.

Focusing on the shopping mall scenario, the most important factors that can disrupt the re-identification process, grouped into the above two mentioned divisions, are explained in detail.

### 1.2.1 Intrinsic Factors that Affect People Appearance

In real situations where people can behave normally with no constrains, people appearance can not be described as a static, time-independent feature. Several actions taken by people, like unusual behaviors, movements or poses, can drastically affect their global appearance seen by a camera in the system. In the case of non-overlapping cameras, this change in the appearance may only affect one camera but not necessary the others, which makes the task of people re-identification very challenging since people appearances may change from cameras.

In the following lines, a list of the main problems that affect the person appearance are described.

- **Person Segmentation** One of the major problems in computer vision is to discard the irrelevant information in an image (normally related to the background) and to focus on the scene elements that are of particular interest (normally addressed as foreground elements). Depending on the scenario and the system requirements, the process of background detection can be more or less complicated. For static cameras, the problem of background subtraction turns to be a relatively easy and closed problem if we assume that the background remains static a considerable part of time. For moving cameras, though, the problem is slightly complicated since it is really difficult to characterise a background that is not static for a minimum time.

  In the majority of situations where people segmentation is required, the scenarios normally represent pedestrian areas with static backgrounds and fixed cameras, in which only pedestrians are supposed to move in the scene. Thus, foreground extraction implicitly involves a person segmentation itself. The accuracy of the person segmentation is highly dependent on the clothes that people is wearing. Therefore, segmentation of people wearing similar colors to the background may be incomplete. Also, people who remain quiet during a considerable period of time would automatically turn into background. Figure 1.2.1 shows several examples of challenging situations. In Figure 1.2.1.a there is a shopping basket between the person and the camera, difficulting the task of segmentation. Figure 1.2.1.b shows a couple that are touching their arms, so that the separation between their segmented silhouettes may not be clear. Figure 1.2.1.c shows a woman carrying several objects, which will degrade the

quality of the segmentation. Finally, Figure 1.2.1.d shows a person wearing black colors as the floor, difficulting the task of segmentation.


(a) (b) (c) (d)

Figure 1.2: These figures show different challenging segmentations. In a) the shopping basket is difficult to be separated from the body. b) shows two people that are touching their arms, so it may not easy to identify the edge of each silhouette. The person segmentation in c) may be difficult because of the carried bag and sweep. Finally, d) shows a man wearing a black shirt, the same color of the floor.

- **Articulated Body Model** The human body has a high number of articulations that permit people to fully interact with the environment and with other humans. This property of the human body produces that people appearance may change depending on the state of all the body joints. Depending on the situation, one can move hands across the head or torso, turn the torso down to pick something laying on the floor or simply walk, where gait produces variations in the global appearance. Figures 1.2.c and 1.2.d show some examples of people that present challenging poses.

  Apart from the appearance changes, the size of the person may also change due to variations in their pose. For example, a person raising a hand or stretching arms out would explicitly change their height and shape.

- **Variation over time** People who are being tracked for a long period of time may probably vary their appearance over time in real scenarios. In idyllic situations, the person appearance would be simply described as a static signature

that characterises the global appearance during all the tracking time. However, people in real situations do not keep a constant gait and are more likely to move hands, stop at certain position, change moving directions, interact with elements in the scene, or even put on or take off complements. These effects produce a temporal variability on their appearance that can be hardly decoupled. Figure 1.3 shows two examples of people whose appearance is remarkably changing over time. Figures 1.3.a and 1.3.b show a person taking a shopping basket from a pile. In this process, the person changes her view regarding the camera, so that in the first frames a frontal view is retrieved, but in the last frames the side and back of the person are captured. Figures 1.3.c and 1.3.d show a person holding a child and taking a shopping basket from the pile. This case is a clear example of how much the person appearance can vary in real situations.


(a) (b) (c) (d)

Figure 1.3: Variation of the person appearance over time. Figures a and b show a woman taking a shopping basket from a pile at the market entrance. Figures c and d show a woman holding a child and also taking a shopping basket.

- **Carried Objects:** Shopping areas are a clear example of challenging scenarios for people re-identification. In these scenarios where people do buy things, people are more likely to grab objects and hold them in their hands. These objects are normally segmented as a part of the person shape since they are attached to the body and are difficult to be excluded from it. Again, carried objects may be seen only by certain cameras, but not for all of them in the

system. When modeling the appearance, these objects can be considered as outliers since they are external objects that do not belong to the person and certainly not represent their appearance. As example, Figure 1.3 showed people carrying objects for a partial period of time of the tracking. Also, Figure 1.4.a shows a person holding a box at the cash desk in a shop. Figure 1.4.b shows a person holding a snacks bag in one camera, modifying thus the person appearance.


(a) (b)

Figure 1.4: Person appearance from two different cameras with the presence of carried objects.

- **Inter-Person Similarity:** A fundamental problem in re-identification is to be able to distinguish among people who are dressed similarly. In crowded scenarios, the probability of having two people dressing similarly is remarkably high. People are normally affected by fashion trends, so it is easy to find places where the most part of the people are wearing the same dominant colors. Besides, flat-textured clothes in black color are the most used by people and also the most challenging. Figure 1.7 shows an example where up to 4 people are wearing jeans combined with black shirts, sweaters or jackets. When these problems happen, one may need to use other information apart from the color. The use of shape information, including height and size measures, or contextual cues, would definitely help in finding the correct match.

### 1.2.2 Extrinsic Factors that Affect the Perceived Appearance

These factors are caused by external circumstances that do not depend on the people themselves and that may affect how they are seen at each camera. Next we give detail about the most common extrinsic factors that may occur in real and challenging scenarios of our scope.

- **Viewpoint:** Imagine a scenario where several cameras are focusing on a single person at the same instant but from different perspectives, overlapping their fields of view. Depending on the position and the perspective of each camera relative to the individual, the person appearance may differ. In general, the

(a) (b)

(c) (d)

Figure 1.5: Example of four different people wearing similar clothes.

appearance seen from frontal, rear or side views may be different in each case since clothing can vary in pattern or colour from each side, or even have details that are only visible from a particular angle such as ties or other complements. On other hand, gait information is also different depending on the point of view of each camera.

Figure 1.7 depicts a situation where a person is seen from two different viewpoints, covering the side/back and front of the individual respectively. Due to the changes in viewpoint, which affect the estimated appearance at each view, the person seen from his back can easily be mistaken for somebody wearing clothes with similar colour and texture. However, these situations are also a challenge for humans if we do not have other contextual information.

- **Intrinsic camera parameters:** One important issue in the image acquisition is the quality and properties of lens and sensors of each camera in the network. Lens do not always behave identically even if they have the same parameters or they come from the same manufacturer. Different color perception and different distortion of the image may be found due to the differences in the focal length or the former materials of the lens. On other hand, camera sensors may also behave differently depending on their technology, CCD or CMOS as an example, or the size of the matrix sensor.

  Depending on the purpose of the camera network and the location of the

(a) (b)

Figure 1.6: The image shows a person seen at the same instant by two different cameras located at different places. Person appearance may be totally different depending on the perspective that they are seen.

recording points, one may need to select different lens and sensors to be able to get the area of interest. Depending on the camera network configuration, the same person could slightly change their pose and appearance among cameras. In order to avoid these problems, it would be recommended to calibrate intrinsic parameters of the cameras in the system.

- **Illumination:** Lightning conditions in uncontrolled environments are mostly unpredictible for both indoor and outdoor locations. Lightning variations directly affect to the color appearance of the person. These variations do not only take part for different cameras, but also for the same camera at different time, depending weather the camera is placed outdoor (where natural phenomenon like day, night, clouds or shadows can affect) or indoor (where artificial lightning may be different depending on the luminance model, shadows...).

  Apart from the changes in the lightning conditions, modern cameras can also adapt automatically to these changes, which result in turn into another factor that needs to be considered.

  Figures 1.7.a and 1.7.b show an example of people affected by different lightning conditions. In the image of the man with the yellow shirt, we can see that in one view his shoulder has been overexposed due to an excess of light, presenting a white colour, where it should actually be yellow.

  Besides, changes in brightness may also occur during the person tracking in a particular camera, as depicted in Figure 1.7.c and Figure 1.7.d.

- **Occlusion:** Occlusions are generated when some part of the person is not visible because there is an element in the scene, let us say another person or an object, which is between the person and the camera. In real and crowded environments, people are likely to be partially or totally occluded by other elements in the scene. This effect may produce severe problems for tracking


Figure 1.7: Two examples of brightness variation. a) and b) show a person wearing a yellow shirt seen from two different cameras. The color perceived by the two cameras significantly changes due to color saturation on the shoulders of the man in b). Figures c and d show a person seen from the same camera in different instants. It can be seen that the scene illumination progressively changes during the person tracking due to the auto-gain property of the camera.

but specially for appearance description. Ideally, when dealing with person description, the area that is occluded should be considered as missing data in the appearance model, so that it is not considered later for matching appearance vectors. In practice, though, one does not know what parts of the person have been occluded. Because of that, the appearance descriptor may be corrupted with the information of the occluding element, which may severely degrade the appearance vector and worsen the re-identification score.

Figure 4.4 shows several examples of occlusion caused by a person and other scene elements. If the segmentation algorithm fails to separate the occluding element from the person of interest, the appearance vector of this person may be seriously corrupted, and also the consequent matching.

- **Contextual Cues:** Apart from the information extracted from the person appearance, one may also want to make use of other higher level information inherent in the scene to generate hypothesis about the person trajectory. In an ideal system, one can suppose that every person who enters into the scene through a camera has to leave the scene by passing again through another


(a) (b)

Figure 1.8: Examples of partial occlusions. a) shows a man occluded by the woman dressed in red colors. b) shows a man almost 100% occluded by the elements in the cash desk. Also, this subfigure shows a woman with a black coat who is partially occluding a man with orange clothes.

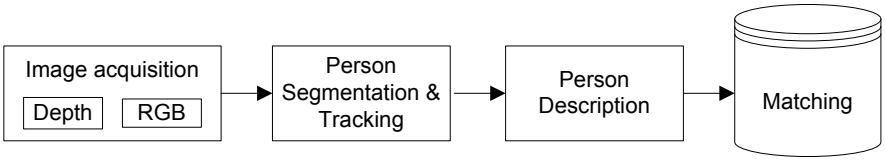



Figure 1.9: System overview

camera. For small shopping malls or shops where this hypothesis may be applied, exploiting this contextual information permits us to establish a prior knowledge of a person at each camera. One can estimate the position of a person in a new camera by using tracking information and time prediction. Also, one can suppose that a person inside a shop will not be all the day inside, so depending on the scenario and the size of the shop, a temporal limit can be set. However, in wider scenarios this hypothesis is not feasible since there may be several entry and exit points outside of the camera view, and too many different paths that a person can follow when he or she leaves a camera.

## 1.3 Overview Of The Approach

In this Thesis a person re-identification system has been proposed to address the problems mentioned above. The general scheme of the proposed approach is depicted in Figure 1.9. The figure shows a generic representation of the main blocks of the system. The first block represents the image acquisition process, which provides a pair-wise set of rgb and depth images. The second block shows the person segmentation and tracking process, which generates a raw descriptor of the person appearance and shape for each frame. When a person that is being tracked abandons the area of interest, the tracking process finishes and the third module

extracts a higher-order descriptor based on the raw information. In module four, this descriptor is matched against the other descriptors in the database.

Because of the challenges described in Sec. 1.2, the task of people re-identification is not an easy and closed problem. The keypoint towards the solution of this problem is to have an accurate representation of the person appearance that can describe a person independently where and when they have been seen. In order to generate a faithful description of the person, an accurate segmentation is crucial. To that end, kinect cameras have been used. To our knowledge, this is the first system used for people re-identification that uses this technology. The key point of these cameras is that they automatically provide accurate depth images synchronized with color images up to 30 fps. The use of depth information eases the task of accurate person segmentation and permits to deal with the real size of the person, regardless of the viewpoint of the camera. This fact allows the system to accurately register the distribution of colors of each person in a 3D coordinate system. Such an accurate 3D color registration is taken frame to frame and forms the low-level representation of a person over time.

The low-level representation of a person is a high-dimensional raw feature vector that contains redundant information in the majority of cases. Depending on the situation, the temporal domain may contain redundant information, or the 3D distribution of color may be not significant in those cases that the person is wearing flat-color cloths. In order to summarize, flatten and strengthen this information, higher-order descriptors are applied over the raw features. As will be discussed in Chapter 4, these descriptors attempt to marginalize either the temporal domain, the angular domain, or both.

The matching strategy that has been followed is based on a distance minimization strategy, where each of the probe person feature vector is matched with the others in the database. A special treatment is carried out for the case of which there is not marginalization in time, where matching people by exhaustive search is not possible given the immensely large search space. Therefore, in this case a class histogram representation has been used into a probabilistic framework to find the class matches, where each class represents a peson in all of their possible appearances.

In order to be able to perform the methods evaluation, we created a public dataset for person re-identification using depth cameras. This dataset contains two scenarios covering different complexities. One scenario gathers people walking straight along a corridor in a school hall. The other gathers people in a commercial establishment with challenging conditions. Further details can be found in Appendix A.

## 1.4 Summary of Contributions

The main contributions of this Thesis are the following:

- Definition of a new surveillance network topology that uses rgb-depth cameras. Description and development of a protocol for managing data.

- Development of an accurate person segmentation algorithm using depth information from rgb-depth cameras. Depth is transformed into a virtual cenital representation, called Height Map, where the segmentation and tracking can be effectively performed. The generation of the Height Maps requires previous calibration to detect the ground plane and to extract the extrinsic parameters of the camera in relation to the floor.

- Development of an algorithm for mitigating the brightness changes in different scenes. The algorithm removes the global variations while preserving the local patterns of color distribution.

- Creation of the first database containing rgb-depth image sequences of people behaving in different scenarios. Two different scenarios have been proposed: a supermarket including 150 different people, which is a challenging environment because of the lightning conditions and the fact that people can carry objects, and a school hall, which is a simpler case with more controlled elements. Databases are publicly available and contain frame-to-frame rgb-depth images together with the groundtruth, which is represented as a binary image where each person is represented by a segmentation mask where all the inner pixels have a common global id that identifies them.

- Development of an algorithm to extract a low-level person features over time. This algorithm generates frame-to-frame, raw information of the person appearance over time. This raw information is generated by quantifying the color in a 3D space using cylindrical coordinates, where the center of coordinates is placed at the center of mass of the pointcloud representing the person at a particular frame. For each bin of the quantified space, the mean color of a 3D patch is stored together with the number of pixels and the mean radius of those elements that contributed into that bin.

- Development of the Bodyprint descriptor, a novelle, compact and robust representation of the person appearance. The descriptor is based on the low-level representation of a person. It marginalises the raw features with angle, radius and time, only preserving the vertical component. Thus, the descriptor is represented by a unique vector which represents the vertical appearance of the person where each component of the vector is a rgb triplet representing the mean values over time and space.

- Development of a cylindrical Bodyprint descriptor that maps the quantified 3D color information of the low-level person segmentation into a 3D cylindrical grid with a fixed radius, which is centered on the center of mass of the person. In this mapping process, the algorithm selects a suitable number of vertical and angular bins according to the scenario requirements. The descriptor calculates the average appearance of the person over time, preserving the cylindrical representation. Note that the cylindrical grid that is used for mapping is aligned at each frame according to the moving direction. This method has the

property of automatically describing the full 3D appearance of a person as the person moves along the camera field of view.

- Development of the novel concept of Latent Feature Models, which are probabilistic dimensionality reduction techniques that provide robust representation of the appearance vectors. These models, which are applied over the Bodyprints, remove strong correlations in the feature vectors, minimise the noise and outliers, and can satisfactorily cope with missing data produced by occlusions.

- Development of the novel concept of Bag of Appearances (BoA) to describe the person appearance. A BoA is a container of color features that fully represents a person by collecting all their different appearances over time. These temporal appearances are marginalisations with angle and radius of the low-level person representation features, but preserving the temporal component to describe each person.

- Development of different matching algorithms for people re-identification that exploit the nature of the proposed features. Bodyprints, cylindrical Bodyprints and latent Bodyprints follow a distance minimization strategy to match the probe signature with the others in the database. Matching of bags is performed in a probabilistic framework by accumulating the probability of matching for all of the elements of each bag with the others in the database.

## 1.5 Outline of the Dissertation

The remaining chapters of this Thesis are logically organized according to the overview illustrated in the Figure 1.9. Chapter 3 summarizes the segmentation and tracking process, together with the low-level person description frame-to-frame. Chapter 4 introduces and assesses different novel person description methods. Section 4.4 introduces the Bodyprint descriptor. Section 4.5 addresses the Cylindrical Bodyprints. In Section 4.6, the novel concept of Bag of Appearances for people matching is introduced and discussed. Section 4.7 addresses the concept of latent features extracted over the Bodyprints. Finally, Chapter 5 summarizes and discusses the main achievements of this Thesis. In this chapter, the future research lines are also addressed.

Chapter 2

# State of the Art on People Re-Identification

Contents



In this chapter, a review of the state-of-the-art techniques for people re-identification is presented. The content in this chapter is organised to describe in a logical order all the steps required in the process of people re-identification.

First, a discussion of the different topologies of camera networks is presented in Section 2.1. Depending on the nature of the problem and the purpose of the application, a different network topology would be chosen. Second, a brief review of segmentation algorithms is presented in Section 2.2. Although tracking techniques are also a key aspect in a general people re-identification problem, they are not addressed in this chapter since our approach does not strongly rely on tracking to describe and match people. A complete review of techniques for person description is presented in Section 2.3. The analysis of the techniques is carried out by dividing the problem according to the nature of the features. Section 2.4 addresses the module of dimensionality feature reduction. Finally, Section 2.5 describes the most used matching techniques for people matching.

## 2.1 Camera Networks Topologies

Depending on the properties and limitations of the scenario where to deploy the re-identification system, one may design a different network architecture in order to provide solution to the scenario restrictions and to be able to cover the area of interest at the necessary image resolution. Thus, the network configuration may

vary depending on the size of the scenario, if there is an open or close area, or weather the system has to work in indoor or outdoor scenarios, etc.

The most common classification is related to the number of observations of a person that the network can compile at each time. According to this classification, the systems can be divided into single-shot and multiple-shot. The former use only one image to perform the person identifiation [Wang 2007], whereas the later use different images of the same person obtained by tracking or by other cameras at the same instant [Bazzani 2010]. The advantage of the single-shot systems rely on the simplicity and ease of use of the gathered information, where each person is identified by a single image at a particular instant. However, re-identification success strongly depends on how representative the image of the person is. On other hand, multiple-shot methods can collect more images of a person, so that the most representative features of the person may be satisfactory covered.

Other classification of networks can be done according to the overlapped area among cameras. We can distinguish between overlapping [Gandhi T. 2006] and non-overlapping camera systems [Wang 2007]. The overlapping-camera systems usually permit to have higher re-identification rates because of the geometrical and spatial cues that simplify the identification and matching reasoning. However, the overlapping topology is not usually feasible in real scenarios because of the high cost that would require this approach to cover extense areas. For this reason, the non-overlapping systems are the most used.

The last representative classification of camera networks has to do with the camera resolution. Depending on the position and perspective of the camera relative to the area of interest, the system can gather more or less people in the field of view. Depending on the size of the person in the image, cameras in the system would need higher resolution than others. The work presented by Tapaswi et al. [Tapaswi 2012] addresses the problem of re-identification in TV series by performing facial analysis. Note that these images have enough resolution to allow the facial analysis. In the work of Bazzani et al. [Bazzani 2010] they use the whole body of a person to compute the descriptor. In their images, people are represented with a good resolution that permits the extraction of accurate global appearance. Other authors [Baltieri 2011] perform people re-identification using smaller representations of people. In this case, the authors base the re-identification on tracking and also using contextual cues to predict the position of person among the cameras in the system.

## 2.2 Person Segmentation Algorithms

Person segmentation is an important task that allows to accurately describe the person appearance in variable scenarios. For fixed cameras, the problem of person segmentation is usually tackled as a problem of foreground object detection, under the common assumption that only people can be considered as possible moving objects in the scene. As stated in the work of Bouwmans et al. [Bouwmans 2008], background subtraction techniques have to cope with several challenges such as

gradual illumination changes, sudden illumination changes, dynamic background, shadows, bootstrapping or signal noise.

There are several approaches to address the problem of foreground object segmentation. In general, the problem basically relies on hypothesis that background can be considered as a fixed element in the scene, and the foreground objects are moving in front of it. The problem is normally addressed by classifying each pixel on the image as either belonging to the foreground or background. The main difference among most of the background subtraction methods is how the background is modeled and updated over time, as explained in [Benezeth 2008].

The basic way to model the background is following the approach of Basic Motion Detection using gray or color images. This approach uses the technique of sliding window with a temporal median filter [Zhou 2001] to characterise the background. Foreground extraction is performed by thresholding a distance metric between the background and current images.

Other strategies model each pixel as a probability density function that has been learned over time. As an example, Wren et al. [Wren 1997] use Gaussian distributions to model the pixel noise at each single pixel. On other hand, Stauffer and Grimson [Stauffer 1999] use multimodal probability density functions to cope with backgrounds containing animated textures such as trees with moving branches, waves on the water, etc. In their work, every pixel is modelled as a mixture of K Gaussians. Other models, such as Kernel Density Estimation, has been proposed in the literature. Elgammal et al. [Elgammal 2000] proposed a Parzen-window estimate of every background pixel to model a multimodal probability density function under an unstructed approach. On other hand, Haritaoglu et al. [Haritaoglu 2000] propose the strategy of minimum, maximum and maximum inter-frame difference to model background using difference of consecutive frames.

Person detection algorithms have also been widely used in the literature to segment people in images without the need of background subtraction techniques. Dalal and Triggs [Dalal 2005] introduced a description method based on histograms of gradients that detects people on still images. The method is previously trained using SVM [Burges 1998] classifier. Viola et al. [Viola 2005] present a fast method for moving person detection, using Boosting techniques based on Haar-like wavelets and space-time differences. Both works of Dalal and Viola are the most relevant in person detection methods in the literature. However, a lot of research has been carried out based on these works. Just to cite a few, Hernandez-Vela et al. [Hernandez-Vela 2012] perform GrabCut human segmentation methodology that is initialised by a person detection module using HOG. They combine spatial information using Mean Shift clustering with temporal coherence using Gaussian Mixture Models. Recently, Xiao [Xiao 2012] presented a method to identify human basing not only in local features but also in context cues in the neighborhood that provide important constraints for detection. On other hand, Subhransu et al [Subhransu 2013] introduce the concept of additive SVM kernels that approximate non-linear SVM kernels by adding approximate classifiers. This method has been demonstrated to yield significant improvements in accuracy over linear SVM while

preserving the computational time.

With the recent introduction of rgb-depth sensors into market, several authors have been addressing the task of person detection and segmentation in the 3D domain. The most relevant work has been presented by Microsoft Research group and is used in gaming on Xbox 360 using a kinect sensor. This work, which is presented by Shotton et al. [Shotton 2011], proposes a method for segmenting people and estimating the 3D position of their body joints. In their article, authors address mainly the process of body parts labeling, but they do not give detail on how the segmentation is carried out. In their released sdk [Microsoft b], they are able to simultaneously detect up to four different people in the scene. Angelov et al. [Anguelov 2005] segment generic objects from 3D scanned data using spin images and MRF. The algorithm needs to be previously trained using SVM. They demonstrate the algorithm performance using a database containing puppets in different positions and with partial occlusions. Gulshan et al. [Gulshan 2011] do person segmentation following a top-down approach. In the coarse stage, they use HOG descriptors to provide a rough person segmentation derived from the bounding box around the person. In the second stage, they refine the segmentation using local grabcut initialised with the predicted segmentation using HOG. Zhao and Thorpe [Liang 1999] segment possible person silhouettes from depth maps that are fed to a neural network that detects pedestrians. Xu and Fujimora [Xu 2003] also extract body silhouettes similarly as Zhao et al. but using a time-of-flight device Finally, Salas et al. [Salas 2011] use HOG descriptors combined with depth information to provide more precise people detection and tracking.

## 2.3 Description Techniques

Accurately describing the person appearance is a key task for the process of people re-identification. The selection of the suitable features has to be done accordingly to the properties of the scenario and the camera network parameters. Features must be able to unambiguosly describe a person in a camera so that this person can be re-identified again in other cameras. To that end, the feature extraction process has to characterise the global appearance of the person but also mitigate the intrinsic and extrinsic effects among cameras outlined in Chapter 1.

A common assumption in the literature, which has been also considered in this Thesis, is that people are wearing the same clothes in all the views. With this assumption, the re-identification problem focuses on how to describe the appearance of individuals and how to match them among cameras. A description of techniques grouped by affinity is presented next.

### 2.3.1 Holistic and Part-based Techniques

Person description techniques can globally be classified into holistic and part-based. Holistic methods use the full body of the person as the unique input where to extract the feature descriptor. These methods calculate global features of the body

to describe the general person appearance. For example, the early work of Nakajima et al. [Chikahito 2003] uses a holistic representation that is used to simultaneously estimating the person pose and the identity using a previously trained multi-class SVM. Other example of holistic representation that has been very popular for re-identification has been introduced by Javed et al. [Javed 2003] [Javed 2005], where the general people appearance is described by global histograms.

On other hand, part-based methods attempt to overcome the problem of unconstrained poses of a person by extracting different features for each body part, such as legs, trunk or head. According to Doretto et al. [Doretto 2011], part-based modelling can be divided into two groups. In the first group, the person bodyparts are identified and modeled by interest points extracted by an interest operator. Mikolajczyk et al. [Krystian 2005] propose the use of a Hessian affine invariant interest operator to extract keypoints from images. Gheissari et al. [Gheissari 2006] propose to extract keypoints over time from foreground patches. This approach is not stable over time, but it is more informative with respect to color variation than the proposal of Mikolajczyk since it increases the probability of generating true correspondences. Other recent research such as the work presented by Martinel et al. [Martinel 2012] also calculate a discriminative signature by exploiting multiple local features. They also provide a novel signature distance measure by exploiting a body part division approach. The second group attempts to describe the person appearance using part-based modeling via model fitting. This approach establishes a correspondence among different body parts suchas the head, amrs, lengs and torso. This is addressed by following a top-down segmentation of the person to localise their different body parts. The most common technique is the decomposable triangulated graph, presented by Amit et al. [Yali 1996] or Doretto et Soatto [Jackson 2008]. These features are matched using model fitting as stated in [Doretto 2011]. This approach has shown good results in comparison to many holistic approaches. Although part-based methods are very promising, holistic methods are still the most followed approach for challenging scenarios.

Passive biometric techniques such as face, iris or gait recognition have been also proposed for re-identification in the literature. Face and iris analysis can be thought as a part-based problem, whereas gait analysis can be included into the holistic group. Despite the high description level that these techniques provide, the low resolution of the images, occlusions and the variety of poses make biometric-based techniques ineffective in many scenarios. One of the few approaches that uses face analysis for re-identification is presented in [Bäuml 2010]. In their work, the cameras are set up to cover a narrow area in corridors. This particular configuration allows increasing the resolution of faces and extracting face tracks that can be matched against a database. Another example that uses faces is presented in [Chen 2007]. In their work, the face modality, if available, is obtained with other global appearance modalities; however, the face modality is only available in 5% of the matches. Gait has been also proposed by several authors as a promising cue for people re-identification [Chen 2007] [Zhang Z 2005] [Bouchrika 2009]. In this case, the difficulty of extracting reliable gait features in unconstrained multi-camera

scenarios may explain why this modality has not been widely used for people re-identification.

### 2.3.2 Color-based Techniques

Color information provides important cues about people appearance since people may wear a wide variety of clothes and colors. Color features are indicated for these cases of low-resolution images since the dominant colors of the people are invariable to the camera resolution. Although people may normally be affected by fashion trends, so that a common color can be predominant in a scenario, or people can wear flat-textured clothes in black color, which do not provide rich information to separate individuals, color-based techniques are the most prominent techniques used for people re-identification.

The main problem of the color-based techniques is the variance with the illumination. Rapid changes in illumination, shadows or differences between indoor and outdoor lightning (color temperature) may drastically affect the color information. In order to remove the color dependence with brightness changes, many researchers have addressed the color constancy problem by searching the suitable color space to better represent the color information and the optimal normalisation of the color features describing a person.

#### 2.3.2.1 Color Spaces

Many color spaces have been proposed in the literature to describe people global appearance. Standard RGB color space has been widely used by several authors [Prosser B. 2010] [Javed 2008] [Cheng 2009] [Bazzani 2010] [Alahi 2008] [Gilbert 2006] [Prosser 2008]. Hahnel et al. [Hahnel 2004] proposed a variation of RGB color space where the RGB channels were combined with luminance so as to remove the intensity information. They only kept R and G channels since B is dependent after normalization. Their results showed that normalising the channels with the intensity information does not help in the re-identification. Wang et al. [Wang 2006] proposed the use of RGI triplet to characterise the color information. This approach improved the one proposed by Hahnel since in this case the global intensity information has been preserved. However, these experiments were only conducted in a single camera, giving no further evidence of the validity of this approach with multiple cameras with different global illuminations.

YUV color spaces have been also proposed in the literature. Jeong et al. [Jeong 2008] use U and V channels to extract a Gaussian Mixture Model applied to the most relevant color clusters, where the centers of the clusters are used as the descriptors. Authors claim that their approach improves RGB-based and YUV-based despcriptors in terms of space and time complexities as well as matching performance. However, they are in contradiction with the early work of Orwell et al. [Orwell 1999] who demonstrated that the use of YUV outperforms the UV using multiple cameras, which leads us to understand that the use of the brightness

information may be considered in the person description.

HSV color spaces have also been used by many authors for people re-identification [Alahi 2008] [hsun Chang 2001] [Kettnaker 1999] [Oliveira 2009] [Gheissari 2006] [Farenzena 2010]. As a representative example, Alahi et al. [Alahi 2008] compared HSV to RGB color space, where they claimed that the use of RGB space slightly outperforms the HSV in similar conditions.

#### 2.3.2.2 Color Normalization

Person re-identification in real scenarios normally suffer from unknown local and global variations of illumination among cameras. In order to minimise the effect of brightness changes, further color processing needs to be performed. The most used method for color normalisation is the color constancy, as explained in the work of Michael et al. [Swain 1991]. They normalise RGB features taken from an object seen from two different cameras. Monari et al. [Monari 2012] also follow a color constancy approach to automatically estimatimate and compensate the local illumination in the scene based on shadow-based illumination maps. Recently, Madden et al. [Madden 2007] introduced the concept of cumulative intensity transformation to create and update k-means color clustering by an adaptative intensity transformation over time. Other authors [Porikli 2003] propose to model the illumination changes among cameras by calculating the correlation between two color histograms of different cameras. Gilbert et al. [Gilbert 2006] extend this work by using an online method to learn the inter-camera illumination changes using RGB color space. However, this methodology implies a training procedure to calculate the correlation among cameras which requires substantial user dedication.

#### 2.3.2.3 Color-based descriptors

Person re-identification based on color-based descriptors has been widely used for many authors since the color information is independent from the camera view and camera resolution, in absense of illumination changes.

The global appearance of a person has been usually addressed using color descriptors. Many authors [Gray 2008] [Javed 2008] [Oliveira 2009] [Farenzena 2010] [Prosser B. 2010] [Kuo 2010] [Zheng 2011] use color histograms for globally describing people, which are easy to be extracted and are scale invariant. Bazzani et al. [Bazzani 2010] use color global histogram combined with recurrent local patterns that characterise texture, after epitomic analysis. The main problem associated with color histograms is that spatial information is discarded. In order to avoid this problem, Dikmen et al. [Dikmen 2010] divide the image into a grid of fixed cells where color histograms are computed at each cell. Other authors [Bird 2005] divide the image into horizontal stripes and characterise color features for each stripe. More recently, Bak et al [Bak 2011] used Mean Riemannian covariance patches for describing feature distributions, considering temporal information of appearance. Mazzon et al [Mazzon 2012] use a centered patch in the upper body

to describe people color appearance. They complement appearance information with contextual data in order to filter people according to potential paths they can follow.

### 2.3.3 Texture-based Techniques

Texture-based techniques for describing the people appearance have been commonly used in the literature to describe global and local appearance of people. Texture information is usually retrieved using well-known local descriptors such as SURF [Oliveira 2009], SIFT [Jungling K. 2011] [Vandergheynst 2009] and Haar [Bak 2010a]. The recent work of Baeuml et al. [Baeuml 2011] present a comparative study of local features for the task of person re-identification where they compare GLOH, SIFT, SURF, HAAR and Shape Context in a common framework.

Part-based approaches based on interest point operators cited in Section 2.3.1 have been tipically addressed using texture descriptors. Hamdoun et al. [Ham 2008] use SURF algorithm to extract and match interest points collected during video sequeces. Cai et al. [Cai 2008] collect a set of patches along the person edges to identify the person, which are matched later using geometric constrints. Oliveira et al. [Oliveira 2009] augment the SURF descriptor using explicit color information to improve matching. Bak et al. [Bak 2010b] train HOG detectors [Dalal 2005] to identify body parts that are described using the region covariance descriptor presented by Tuzel et al. [Oncel 2006]. Farenzena et al. [Farenzena 2010] also use texture information by searching recurrent local motifs with high entropy in order to divide the body into three parts, which is later represented by color histograms and maximally stable color region descriptors as explained in the work of Forssen et al. [Forssen 2007]. Bak et al. [Bak 2010a] use haar-like features and dominant color descriptor (DCD) clustering to obtain invariant and discriminative signatures.

In other works, texture information is concatenated with color information into a long descriptor as in [Tao 2015], where LBP features are merged with RGB and HSV color histograms. Another alternative for fusing color and texture information is proposed in [Lan 2014], where LBP features based on the quaternionic representation are used.

### 2.3.4 Shape-based Techniques

In this Thesis, the concept of shape information has been treated as the 3D information retrieved from a person body. Generally speaking, descriptors for 3D shapes have become popular during the recent years due to the advances in modelling, digitizing and visualizing techniques for 3d objects. The problem of 3D object matching has been commonly addressed as a shape matching problem, where color and texture information are discarded [Zhang 2007].

The basis of shape-based matching relies in the concept 3D descriptors. In general, the process of 3D object description can be considered as a mapping function that transforms from high-dimensional shape features to a lower-dimensional space

in which the most relevant information is preserved and data can be efficiently and effectively processed.

Many 3D shape descriptors have been reported during the last decade, as described in the surveys of Zang et al. [Zhang 2007] and Bustos et al. [Bustos 2005]. Shape descriptors can generally be divided into three main groups. One group includes all feature-based descriptors, where features can be subdivided into four subcategories according to the type of shape: global features, distribution-based features, spatial maps and local features. The first three subcategories represent an object by a single global shape descriptor vector representing its overall shape. Some examples of these descriptors are Shape Histograms [Ankerst 1999], which are an example of distribution-based descriptor or Extended Gaussian Images [Duda 2007], which are histogram-based. Although these approaches are the most common in the literature, global features normally fail to capture the specific details of the object. For that reason, local-based features, which are another subcategory of feature-based descriptors, collect a set of descriptors centered on some interest points on the surface of the object. As an example, Radial-cosine transform [Zaharia 2001] are local-based feature descriptors and 3D Shape Contexts [Kortgen 2003] are semi-local feature descriptors. Note that these methods capture major details of the objects, but may generate too unfeasible and heavy descriptors.

The second classification includes all the graph-based descriptors. The concept of description is totally different from the vector-based descriptors mentioned above. Graph-based descriptors are more complex and more difficult to be calculated, but they provide geometrical and topological shape information of the objects, which means that these descriptors can represent relationships among parts of the object. However, they are not valid for general purposes since they require dedicated measures and matching schemes. Some examples of graph-based descriptors are the Multi-resolution Reeb graphs [Hilaga 2001] [Tung 2005], which are have the potential of encoding geometrical and topplogical shape properties, but are not applyable to all kind of shapes since they rely on the selection of the appropiate Reeb function [Lyer 2005]. Another example is the Skeletal graphs [Sundar 2003], which are Direct Acyclic graphs (DAG) that describe geometric features and allow parts matching.

The third group includes some heterogeneous descriptors such as complex EIG [Kang 1993] or 3D Zernike moments [Novotni 2003] just to cite a few.

Focusing on the particular problem of person re-identification, only a few authors have addressed this problem using 2.5 or 3D shape information. This fact is because the 3D information obtained in a real scenario for people re-identification is not the same accurate as in the case of static object recognition, where their 3D shape is retrieved in optimal conditions. The concept of using 3D information for person description is relatively new in the field of person re-identification and is becoming more popular with the recent introduction of rgb-depth cameras that have brought the opportunity to easily extract accurate shape information from people. Note that the above mentioned methods only use shape information for describing and matching objects. In the case of person re-identification, authors normally use color information together with shape information to improve results and compen-

sate the bad accuracy in the estimation of the depth information. For example, Gandi and Trivedy [Gandhi 2007] use histograms to characterise people with the so-called panoramic appearance map for re-identification in camera networks with overlapping field of views. They use the geometry of the camera network to generate a panoramic map centered on the person's location, and they use temporal information to register multiple maps over time, improving their quality. Papadakis et al. [Papadakis 2010] introduce a cylindrical 3D descriptor for generic object re-identification in controlled scenarios. The cylinder is filled up with the projection of a set of 2d panoramic views of the object. Recently, Baltieri et al [Baltieri 2011] proposed a 3D generic rigid body model that is filled up using person appearance acquired with 2D calibrated cameras. The estimation of the person height from a calibrated camera allows to map silhouette points from a 2D view to the 3D rigid model, which they call sarcofag. The model is scaled according to person height, but no shape information is actually retrieved from the person. Finally, the work of Salas et al. [Salas 2011] present a methodology that combines color and depth information from rgb-depth cameras to perform re-identification in indoor scenarios. They combine the output of a Histogram of Oriented Gradient (HOG) person detector with the tracking information generated in 3D domain to identify each person

## 2.4 Dimensionality Reduction Techniques

Dimensionality reduction techniques have been widely used in machine learning for the purposes of data visualization, data compression, noise removal, pattern recognition, exploratory analysis and time series prediction. In this work, dimensionality reduction has been performed in the field of pattern recognition to further collect the relevant data of the descriptor vector in a lower feature space (where the features representing the object may provide more representative information), while removing strong data correlation, minimising the noise of the measures and dealing with missing data vectors. Depending on the nature of the observations, techniques have been traditionally classified [Fodor 2002] into linear methods such as Principal Component Analysis (PCA), Factor Analysis (FA) or Projection Pursuit (PP), and non-linear such as Independent Component Analysis (ICA) or non-linear PCA, just to cite a few.

As an example, Yang et al. [Yang 2011] perform person re-identification using an eigenspace appearance representation. Appearance is firstly described using color and spatial information at each pixel. Given that the person appearance may vary over time, they take some keyframes of the person and apply kernel PCA to represent the manifold. The similarity measurement is taken in the reduced space. Prosser et al. [Prosser B. 2010] address the problem of person re-identification as a ranking problem, where they learn a subspace where the most likely match is provided as the highest ranking. They treat the problem as a relative ranking problem, instead of an absolute score problem. Satta et al. [Satta 2011] [Satta 2012] proposed a new tech-

nique for dimensionality reduction applied to appearance descriptors in a proposed unifying framework that they name Multiple Component Matching (MCM). They propose a methodology for transposing appearance descriptors in MCM framework into a dissimilarity form that they name Multiple Component Dissimilarity (MCD) framework. This framework compacts information and speeds the process of comparison, although re-identification accuracy has been demonstrated to be lower. Pedagadi et al [Pedagadi 2013] combine color moments and color histograms as features. Their work follow a metric learning approach that combines PCA and Local Fisher Discriminant Analysis in two stages to carry out the person re-identification.

Focusing the attention on the problem of this Thesis, a probabilistic treatment of the dimensionality reduction techniques appears to be of particular interest due to the nature of the input features, which may have missing data and high dimensionality. This fact leads to the use of probabilistic approaches with iterative solutions to tackle the problem. A smart solution to the problem is to use the method of probabilistic PCA (PPCA) proposed by Tipping [Tipping 1999] and Roweis [Roweis 1998]. PPCA approach is a good choice since it inherits the qualities of well-known standard PCA, and can be expressed as a probabilistic latent variable model problem. PPCA bases on a linear-Gaussian framework in which all of the marginal and conditional distributions are Gaussian. The probabilistic treatment has the advantage of solving some problems of direct PCA such as missing data due to occlusions and outliers in an elegant way.

Other probabilistic methods, such as Factor Analysis (FA) [Basilevsky 1994], are also linear Gaussian latent variable models similar to PPCA. The only difference between them relies on the covariance matrix of the observed data given latent variables. In PPCA it is full covariance matrix, whereas in FA it is diagonal.

## 2.5 Matching Techniques

The task of person re-identification among different cameras consists in finding a unique correspondence of a query person within a probably huge set of persons in a database. This problem is usually addressed as a matching problem and slightly differs from the problem of person tracking, which benefits from the temporal information for identifying the person over time.

The problem of matching people has usually been addressed using standard distance metrics such as correlation, L1 [Wang 2007] and L2-norms [Ma 2012] [Hu 2006], Mahalanobis [Hirzer 2012] or Bhattacharyya [Comaniciu 2003] [Javed 2003] to enumerate a few. Techniques such as Histogram Intersection [Gheissari 2006] have also been proposed in the literature to compare histograms of color images, which have been widely used due to its property of rotation and scale invariance. Other techniques such as Kullback-Leibler have been used by authors [Yu 2007] [Kang 2004] to extract the similarity between two probability density functions representing the persons appearances. As a summary, Alahi et al. [Alahi 2010] collects a complete comparison among these techniques. In their

work, they claim that Bhattacharyya metric has slightly better performance than the other metrics.

The use of complex appearance descriptors may require the use of tailored matching metrics. For example, Farenzena et al. [Farenzena 2010] propose a matching method named SDALF ( Symmetry-Driven Accumulation of Local Features) which can group heterogeneous features gathered from a person over time. Authors claim that their method is simple and effective and is independent from the number of candidates in the database. Madden et al. [Madden 2007] propose a multi-feature matching context that mixes color information with texture information to identify a person.

In multi-shot networks where cameras can get multiple instances of a person over time, several methods have been presented to exploit temporal information. Zheng et al. [Zheng 2011] address the matching problem as a distance learning problem that seeks for learning the optimal distance for true image pairs. They introduce the model of Probabilistic Relative Distance comparison, which makes the system more robust to appearance changes and overfitting. Bedagkar-Gala et al. [Bedagkar-Gala 2011] propose a part-based spatio-temporal model that gathers the changes of the person appearance over time using two different color features to retrieve the chromatic content and the representative colors. The re-identification is done by solving a linear problem to minimize the total cost between the target descriptor and the gallery.

Other authors address the re-identification as a learning metric problem. Dikmen et al. [Dikmen 2010] uses a learning framework that permits to classify images using nearest neighbor approach, where neighbors beyond a learned distance are rejected. The work of Prosser et al. [Prosser B. 2010] learn a subspace using SVM where features are matched following a ranking classification. On other hand, Hirzer et al. [Hirzer 2012] learn a PCA kernel to project data onto a simpler space where nearest neighbor using Mahalanobis distance is selected for feature matching under the reduced space.

Chapter 3

# Person Segmentation and Tracking

Contents



## 3.1 Introduction

The process of person re-identification in video sequences can be easily divided into three main parts: person segmentation, person description and matching. In this chapter we address the problem of person segmentation and slightly cover the low-level representation of the person appearance and shape that will be used in latter chapters.

Person segmentation is a key aspect in the re-identification process since the quality of the description is directly related to the quality of the segmented person, which is taken as the descriptor input. When the person segmentation task is performed in surveillance scenarios, the cameras are normally considered fixed, the background can be considered static for a relative period of time and only people can move in the scene. Under these common assumptions, the problem of person segmentation can be directly addressed as a background subtraction problem, where

the output of the algorithm represents the people in the scene. In idyllic scenarios, the relation is straightforward. However, in real scenarios there is no certainty about this hypothesis since other elements can take part in the scene. In these cases, the interest objects retrieved by the background subtraction algorithm need to be post-processed in order to identify the people in the scene and filter out other undesired objects. Although it is possible to directly detect the people in a scene with no need of any background subtraction method [Dalal 2005], this approach does not produce clean people segmentation, which directly affects the quality of the person descriptor by adding noise. For this reason, researchers normally use the background subtraction approach under the mentioned assumptions.

Background subtraction has been widely used in the literature for several different fields such as traffic monitoring, video surveillance or human-computer interaction, to enumerate a few. This broad use has prompted the development of many different algorithms in the research community together with many higher-level processing for improving algorithms output. The concept of the background subtraction relies on building an explicit model of the background that is updated over time. The difference among methods mainly relies on how they handle the background update over time, and how they post-process the information to provide compact information about interest objects.

Despite the advances on this topic, several problems still remain open mainly due to the limitations of the conventional cameras. Some of these problems are listed below:

- Stationary people who do not move during a long period of time can be considered as part of the background. Although this situation is not usual, it is a problem that standard methods cannot easily cope with it.
- People wearing cloths with similar colors to the background or the floor can be a problem for low-level background subtraction algorithms, since those body parts with similar colors can be misclassified as background.
- Relatively rapid changes in background may produce outliers in the estimation of the foreground objects. Although algorithms are normally tuned to adapt to background variations, any background change that is produced as rapid as the foreground does, would generate an outlier in the foreground estimation.
- Sudden and rough illumination changes may also affect the estimation of the background.
- Shadows that are cast onto the floor or over objects may affect the estimation.
- People that are too close or touching together may be segmented as a unique person by standard algorithms.
- The fact of simply taking the output of the background module directly as the person implies that other moving objects may be wrongly considered as

a person. This problem may be solved by taking higher order processing to detect people over the interesting objects output by the background module.

- The estimation of the person may be significantly damaged if other element is occluding part of the target person. In this case, the occluding element may be considered as part of the person silhouette, providing wrong information about the real appearance.

## 3.2 Motivation and Contributions

The recent introduction of revolutionary rgb-depth cameras into market has brought a promising future to the field of image processing. The use of depth information can remarkably improve many of the current rbg-based image techniques, while also offering new possibilities to tackle problems in which simple 2D images may fail. Despite this favorable framework for the image processing field, the incipient use of these cameras in the market implies that few work can be found in the literature to address common problems such as the person segmentation task. Therefore, this work has been motivated by the lack of 3d techniques in person segmentation that would be usefull to solve many of the problems mentioned in Section 3.1.

In this chapter, a novel solution for person segmentation using rgb-depth cameras has been developed to provide high accurate person segmentation. In particular, the proposed algorithm transforms the depth map obtained from the camera into a height map, which is a cenital representation of the scene. Using this particular representation, people can be easily extracted by exploiting the shape information of the scene objects. The person segmentation is enterly conducted using depth information, discarding the color. Information of scene elements such as distance to the camera, height or volume provide essential cues for detecting people in the scene. Although the basis of the algorithm is similar as in the case of 2D images in the sense that it relies on the moving objects to identify people, the proposed algorithm provides a solution for many of the problems cited in Section 3.1. For example, stationary people are not anymore a problem in this framework because their height and volume clearly identify them into the scene. All the other color-dependent problems such as people wearing clothes with similar colors to background, changes in illumination or shadows that appeared in the case of 2D framework are not longer a problem in the 3D domain. Also, problems of people touching their bodies can be solved by using 3D information. On other hand, there is no need to use person detector methods to filter out people within the foreground objects, since people can be detected attending on volumetric information such as shape or height. Finally, occluding objects to people can be discarded using depth information, producing missing data at the corresponding occluded parts of the person.

In this chapter, the problem of person tracking is also tackled. Due to the goodness of the 3D domain, accurate person tracking is performed using the nearest neighbor approach to match the segmented people over time.

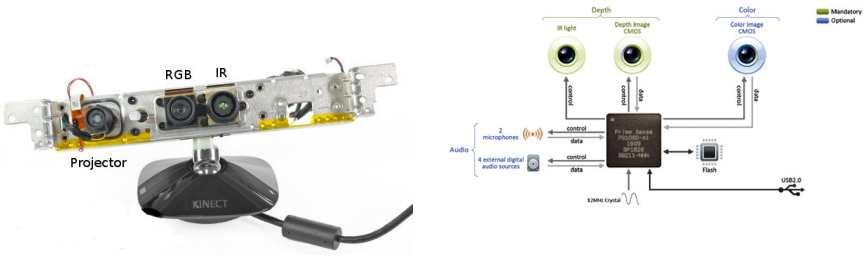



Figure 3.1: Kinect sensors: rgb camera ir camera and ir generator [Microsoft a].

Finally, an algorithm that extracts a low-level person description from the segmentation has been presented. This basic representation of the person is taken as the input for the proposed descriptors. This representation basically discrtetises the information of the 3D domain into a 3D grid for faster processing. The information that is stored at each bin is the mean color, mean radius and the number of pixels that have contributed to that bin.

## 3.3 The RGB-Depth sensor

The kinect sensor was introduced by Microsoft in November 4, 2010 as an input device for its game console Xbox-360, initially under the name "Project Natal". However, the depth sensing technology behind kinect was invented in 2005 by Zalevsky et al. [Zalevsky 2005], from the company PrimeSense [PrimeSense ]. Shortly after kinect commercialization, PrimeSense released a library called OpenNI [Openni ] to ease the development of interactive applications. OpenNI provides an easy API to get access to RGB and depth information, and provides the functions to segment and estimate the person pose. The SDK can segment up to four people and is able to estimate their pose using 48 skeletal points at 30Hz. The SDK also provides the functionality to recognise skeleton actions as well as voice commands. At present, PrimeSense SDK it is used by well-known open source projects such as OpenCV (starting at version 2.2) [OpenCV ], ROS [ROS ] and PCL [Pointcloud ]. After PrimeSense release of the SDK, Microsoft Research released an SDK with identical functionality as OpenNI [Microsoft b] including gaming examples. The last release of Microsoft also permits the automatic initialisation of the skeleton tracking without the need of initial calibration. Note that although these libraries kindly offer person segmentation tools, none of them is valid to be used in the scenarios presented in this Thesis. These depth cameras have been specially designed for gaming, so the camera is relatively close to the people. In the scenarios of this Thesis, the cameras are placed further from the people, which produces significant smaller people sizes that are not compatible with the SDKs requirements.

The keys of success of kinect sensor rely on the fact that the sensor is able to provide RGB and depth images with 640x480px up to 30 fps. The sensor has

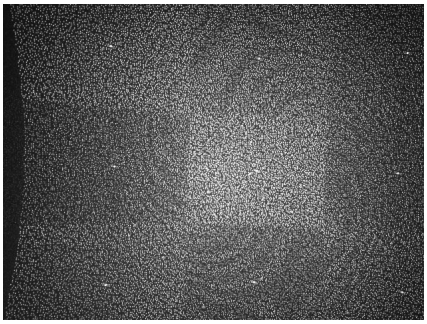


Figure 3.2: IR pattern projected by kinect.

Figure 3.3: Examples of aligned RGB and depth images obtained with kinect.

an internal fpga implementation that accurately extracts the depth from the input sensors, which are: a color RGB camera and an infrared (IR) camera that captures the pattern (see Figure 3.2)projected by an IR pattern generator. Figure 3.1 shows these sensors inside the kinect frame. The device can be configured to *align* both images (RGB and depth). This is required because both cameras have a different focal length and are located at slightly different positions. An example of RGB image and its associated depth is shown in Figure 3.3. For a number of reasons, depth information may not be available for all pixels. In that figure, locations where no depth information is available are shown in black. They are produced by materials with strong reflection (floor) or magnetic components (monitors), where the IR pattern projected by the emitter is not well reflected and therefore cannot be correctly received by the camera sensor. RGB and depth images are aligned taking the RGB as reference. The sensor has a minimum distance to measure depth of around one meter, and a maximum distance of about 10 m. These values do not represent any problem for many indoor surveillance scenarios. However, a limitation of this sensor is that it can only work in indoors environments under normal illumination conditions. In outdoors, the IR pattern has not enough contrast to measure distance.

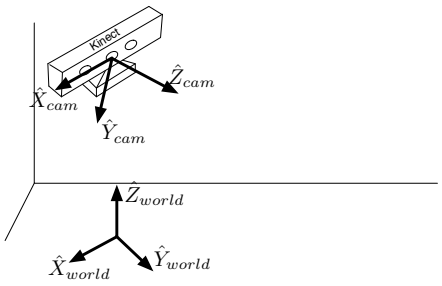



Figure 3.4: Relation between camera and world coordinates.

## 3.4 Scene Calibration

### 3.4.1 Methodology

Camera calibration is the first step towards the person segmentation. The calibration task consists in finding the kinect extrinsic parameters (**R**,**T**) that allows to transform spatial camera coordinates to world coordinates, generating the particular Height Map representation that will be detailed in Section 3.5.

The spatial camera coordinates are given by the depth information and intrinsic kinect parameters. Thus, the $k$-th pixel in a depth image provided by the kinect sensor can be determined using the following equations:

$$\begin{aligned} z_k^{cam} &= d_k \\ x_k^{cam} &= (x_k - \bar{x})\, z_k^{cam}/f \\ y_k^{cam} &= (y_k - \bar{y})\, z_k^{cam}/f \end{aligned}$$

where $f$ is the focal length (in pixels), $x_k$ and $y_k$ are the pixel coordinates in the image (in pixels), $d_k$ is the depth associated with the pixel, and $\bar{x}$ and $\bar{y}$ are the coordinates of the image center. Previous equations provide 3D coordinates in a coordinate system that has the origin at the optical center of the camera, and that it is aligned with the camera axis (see figure 3.4).

The process of camera calibration seeks for a new coordinate system where the plane $\hat{X}\hat{Y}^{world}$) coincides with the ground plane; the $\hat{Z}^{world}$ axis in the world coordinate system is aligned with the true vertical axis; and the origin of the world coordinate system is located on the ground. This way, the $z^{world}$ coordinate represents the height with respect to the ground plane. Figure 3.4 shows the camera and world coordinate systems.

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

$$\begin{pmatrix} x^{world} \\ y^{world} \\ z^{world} \end{pmatrix} = \mathbf{R} \begin{pmatrix} x^{cam} \\ y^{cam} \\ z^{cam} \end{pmatrix} + \mathbf{T} \tag{3.1}$$

Some full automatic methods exist in the literature to determine the ground plane [Hansen 2007]. However, a more practical approach has been used to that end

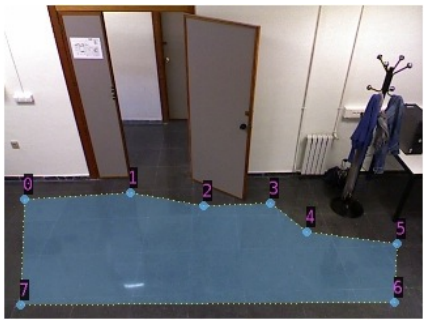



Figure 3.5: Portion of ground used for calibration.

in this Thesis. This approach requires to manually select a portion of the RGB-d image that corresponds to the ground plane (see Figure 3.5). For each point under the mask, the method calculates the camera coordinates of each point for which depth data is available. The ground points are assumed to be on a plane in space, therefore the eigenvector associated to the smallest eigenvalue of the covariance matrix of the ground points will give the direction corresponding to the normal of the plane. Let's call this direction $\hat{Z}^{world}$. At this point, we only know that $\hat{Z}^{world}$ is normal to the ground plane, but we do not know yet if it is pointing upwards or downwards. To resolve this ambiguity, we compute the projection of each ground point to the $\hat{Z}^{world}$ direction:

$$z_k^{world} = \vec{P}_k^{cam} \cdot \hat{Z}^{world} \tag{3.2}$$

where $\vec{P}_k^{cam} = (x_k^{cam}, y_k^{cam}, z_k^{cam})^t$ and '$\cdot$' denotes the dot product. We are interested in getting a unit vector $\hat{Z}^{world}$ that points *upwards*. The mean value of $z_k^{world}$ of the pixels under the ground mask is the relative height of ground with respect to camera. If this value is negative, it means that ground is lower than camera. This is the normal situation as shown in Figure 3.4, and in this case $\hat{Z}^{world}$ is actually pointing upwards. In the opposite case, we just change its direction, $\hat{Z}^{world} \leftarrow -\hat{Z}^{world}$.

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

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

and then normalizing for unit norm:

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

For $\hat{Y}^{world}$, the cross product between $\hat{Z}^{world}$ and $\hat{X}^{world}$ is computed

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

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

$$\mathbf{R} = \begin{pmatrix} \hat{X}^{worldT} \\ \hat{Y}^{worldT} \\ \hat{Z}^{worldT} \end{pmatrix}$$

Finally, we need to determine the components of the translation vector $\mathbf{T}$. The $t_x$ and $t_y$ components are irrelevant for the segmentation process, and we have set them to be zero (the world coordinate origin lies just under the camera as shown in Fig 3.4). The value of $t_z$ has been chosen so that ground points have a height of around zero:

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

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

### 3.4.2 Calibration Results

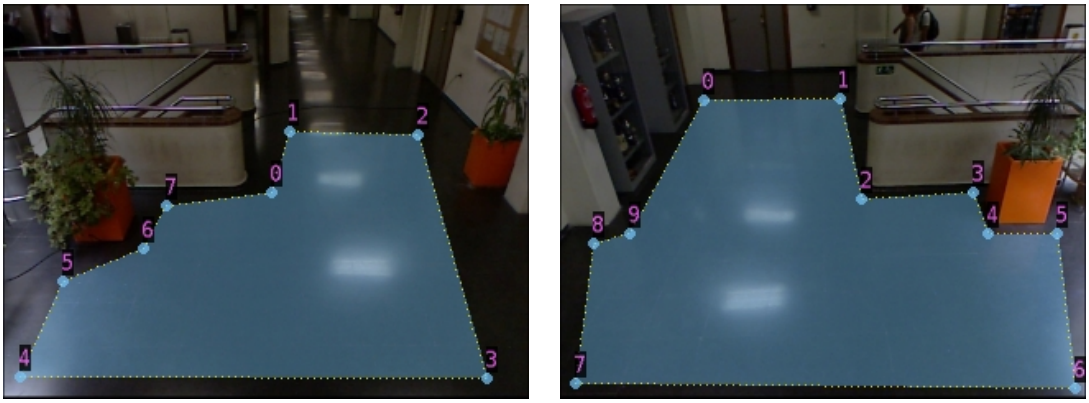
Below are shown the results of the calibration for each particular scenario and for each camera. The camera rotation and translation are shown for the school hall and supermarket scenarios. Cameras are approximately placed at the same height and have similar pitch, which produces similar calibration matrices. These camera poses allow to cover wide areas (up to 6 meters in ground plane) while having sufficient image resolution for the processing. This does not mean that people are seen by the cameras from the same perspective, since cameras may not be covering the same view. In the school hall, each camera gets the front and rear parts of the people in the scene. In the supermarket, one camera covers the side and frontal view, whilst the other only covers the frontal view. However, in the supermarket people move more freely, so people may rotate in the scene and cameras may catch many views of them.

Table 3.1 shows the calibration results for the school dataset. Note that in the estimation of the ground plane, we have previously removed the outliers using Ransac [Fischler 1981], where the algorithm took 22784 points with 20107 inliers for the entrance camera, and 24460 points with 19141 inliers for the exit camera. The quality of the estimation of the ground plane depends on the selected area shown in Figure 3.6. The wider the area is, the more number of points are used for the estimation. Note that it is important to have a considerable area since the floor plane in the depth map may have gaps and inaccurate depth values due to light reflection on the floor tiles.

The calibration results for the supermarket are shown in table 3.2. In this case, the ransac algorithm for filtering outliers on the floor plane has used 14538 points with 13872 inliers for the camera at the entrance, and 16780 points with 14839 inliers. Figure 3.7 shows the roi used to calculate the floor plane. Note that in this

Table 3.1: Calibration results for the school databse.

| | Entrance | Exit |
|---|---|---|
| R | $\begin{pmatrix} 0.99 & -0.03 & 0.03 \\ 0.00 & 0.62 & 0.78 \\ 0.04 & 0.78 & -0.62 \end{pmatrix}$ | $\begin{pmatrix} 0.99 & 0.01 & -0.01 \\ 0 & 0.56 & 0.83 \\ -0.01 & 0.83 & -0.56 \end{pmatrix}$ |
| T | $\begin{bmatrix} 0 & 0 & 3.01 \end{bmatrix}$ | $\begin{bmatrix} 0 & 0 & 2.85 \end{bmatrix}$ |



Entrance Exit

Figure 3.6: Selection of the roi used for the calculation of the ground floor used in the school hall database.

scenario there are less points to estimate the ground plane than in the case of the school hall due to the smaller size of the interaction area.

Table 3.2: Calibration results for the supermarket databse.

| | Entrance | Exit |
|---|---|---|
| R | $\begin{pmatrix} 0.96 & -0.13 & -0.23 \\ 0.26 & 0.41 & 0.87 \\ 0.03 & 0.90 & -0.42 \end{pmatrix}$ | $\begin{pmatrix} 0.99 & -0.01 & -0.12 \\ 0.09 & 0.68 & 0.72 \\ -0.08 & 0.72 & -0.68 \end{pmatrix}$ |
| T | $\begin{bmatrix} 0 & 0 & 2.92 \end{bmatrix}$ | $\begin{bmatrix} 0 & 0 & 2.95 \end{bmatrix}$ |

## 3.5 Height Maps

### 3.5.1 Definition

The scene calibration allows to create the height map representation, which is a virtual cenital view of the scene. A height map can be thought as an image where its pixel values represent the height with respect to the ground. This representation, which has also been used in other works [Zhao 2005], makes the task of segmentation and tracking much simpler.

The process used to obtain height maps starts by representing all pixels, for

Entrance Exit

Figure 3.7: Selection of the roi used for the calculation of the ground floor used in the supermarket database.

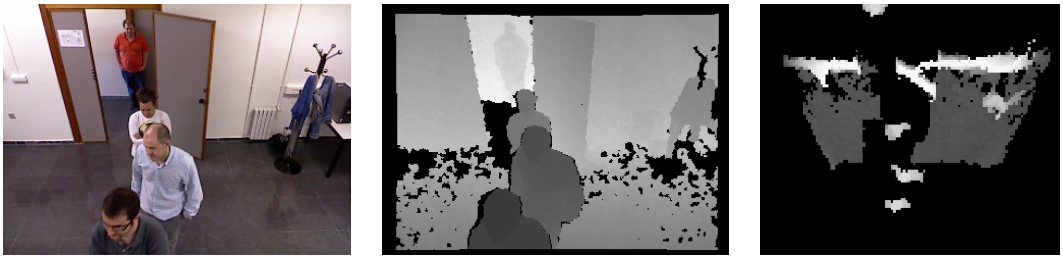


Figure 3.8: Height Map example and corresponding image and depth.

which depth information is available, in world coordinates using the transformation of eq. 3.1. Then, the plane $XY^{world}$ is quantized in small bins of size $5 \times 5$ cm. Pixel $k$ is assigned to the nearest bin according to its $x_k^{world}$ and $y_k^{world}$ coordinates. Finally, each bin is represented with the maximum height (max. $z_k^{world}$) of its assigned points. Figure 3.8 shows an example of height map. The origin (position of the camera) is at the bottom center of the height map. Upwards in the map means getting further from the camera ($\hat{Y}^{world}$ axis). Horizontal axis on the map corresponds to $\hat{X}^{world}$. On the map, black means a bin for which there has been no point falling into it (mostly due to occlusions and the field of view). Dark grey means a height around zero (ground plane) while brighter intensity means higher with respect to the ground.

Using this representation (where the z-axis represents the person height) as the input for the later people description module, makes the description process non-dependent on the camera view angle, which is a common problem in re-identification merely based on rgb image input.

### 3.5.2 Restrictions

Depending on the application purposes and the distribution of the elements of the scenario, one may find interesting to impose certain restrictions on the scene in order to ensure that the re-identification process is done correctly.

In the school scenario, which can be thought as a corridor with only one possible path, users have to necessary pass through the door in a straight movement. The scenario itself imposes a limitation in the people behavior in the scene which eases the process of description and re-identification, ensuring that all the people that come into the scene leaves it from the other camera. However, the supermarket scenario presents a more unconstrained case, where people can move randomly in the scene. Thus, we want to ensure that the system strictly tracks and analyses people that enter into the shop and then leave it. To do that, we need to discard employees that may move in the entrance area and even other clients that have previously entered the shop and consequently, they have already been tracked and described. At the exit, the process is similar. We want to discard people who approach the checkout lanes but finally leave the cue to still buy more things, and also discard employees that work near the checkout lanes.

In order to ensure that people are strictly described once at entrance and exit, we use a set of virtual fences based on the height maps. These fences are used to:

- Mark out the interest areas where people can be analysed. Outside these areas people are not neither tracked nor described.

- Consider only tracks that have entered the area from the entrance door. And similarly with the exit. Thus, we ensure that people who has not come from the entrance door are either employees or already known clients, so they do not need to be described again.

- Ensure that the client finally enters or exits the shop, and does not go back. This is achieved by checking that the final position of the user regarding the initial position.

Figures 3.9 shows the height maps of the supermarket database for the entrance and exit areas, in which color lines indicate the virtual fences. The green line delimits the area of interest where users are going to be segmented and tracked. Outside this area, there is not image analysis. The red line indicates the virtual fence that the client has to cross in order to be considered by the system. The red arrow indicates the direction that the user has to follow in order to be finally considered by the system. If the user movement goes contrary to the arrow direction, the user is not finally considered for comparison.

## 3.6 Person Segmentation and Tracking

### 3.6.1 Segmentation

From the example in Figure 3.8, it can be seen that people appear as clearly separated bright blobs in the height map. The process of segmentation is carried out by thresholding the height map and analysing connected components in order to discriminate people. The following steps describe the process of people segmentation algorithm:

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

2. Threshold the height map at an appropriate level. In this Thesis, the threshold of 120 cm. above the ground plane has been used.

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

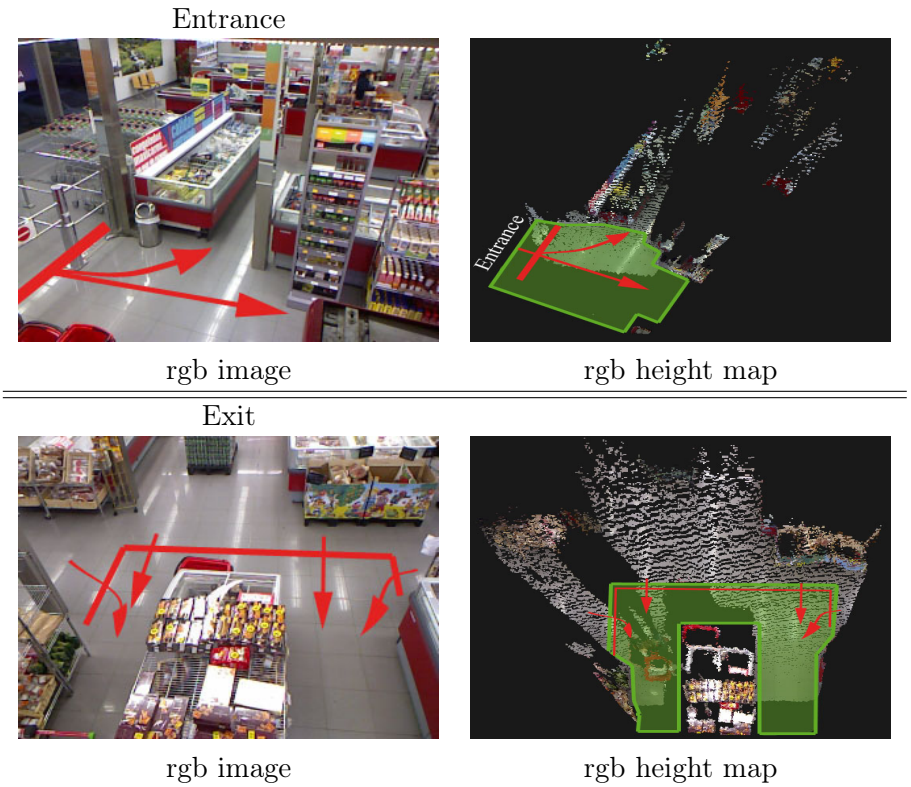



Figure 3.9: Height Maps for the supermarket scenario. The first row shows the rgb image seen from the camera point of view together with the height map at the entrance. The second row sows the same information for the exit. The red lines represent the virtual fences that the user have to cross in the direction pointed by the red arrow in order to be detected. The green line delimits the interest area where the users are analysed. Outside the green area, users are not analysed.

4. Each connected component (blob) may correspond to one or more people. If the blob is large enough we try to split it. To do so we observe that heads are local maxima of the height map. To prevent over-segmentation, the idea of contrast of a local maximum is borrowed from mathematical morphology [Grimaud 1992, Vachier 1995]. We remind that the contrast of a regional maximum is *the minimum descent in order to reach a higher maximum*. Figure 3.10 shows a detail of a blob corresponding to two people where two contrasted maxima are clearly visible. For each connected component determine the contrast of the regional maxima and select those that have a contrast of at least 20 cm.

5. If more than one regional maximum with enough contrast is found within a blob, we will subdivide it using the watershed algorithm using the valid maxima as markers [Beucher 1992].

6. The number of image pixels associated with each local maxima is retrieved and those that do not contain enough image pixels are discarded. Using the proposed databases, this number has been set to 50 pixels. This avoids spurious small objects created by distance noise.

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

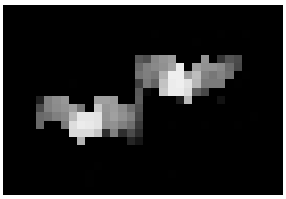


Figure 3.10: Portion of height map showing the presence of two contrasted maxima due to two heads.

An example of good segmentation is depicted in Figure 3.11, where the left figure shows the segmentation result onto the rgb image and the right figure shows the segmentation seen from the height map representation.

A few remarks are pertinent at this point:

- Segmentation provides us information of which pixels in the original image belong to each particular person.
- For each pixel in the original image we can know its height ($z_k^{world}$) and its RGB value.
- An estimate of the height of each person is readily available from the segmentation as the maximum value with a given label in the height map. This can be useful as a first clue in the process of matching people at entry and exit (only people of similar height will be tested).

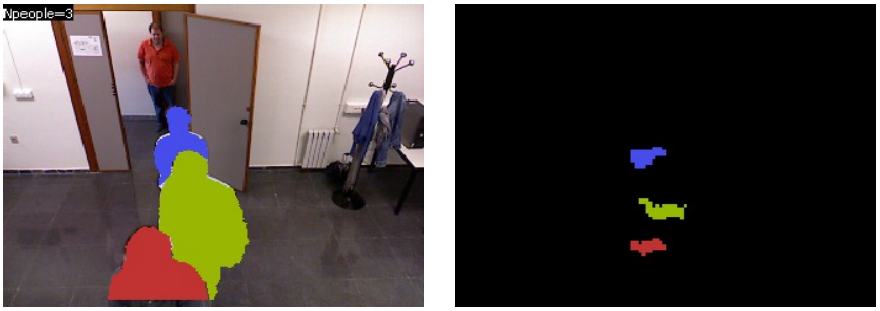



Figure 3.11: Segmentation results of example in Figure 3.8.

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

### 3.6.2 Tracking

Tracking is the process of linking the segmentation results from several frames over time. Henceforth, a *track* will denote a thread of linked objects corresponding to the same person at different instants. Since quality of the segmentation is very good and the frame rate is relatively high compared to the motion of people, the proposed tracking strategy is quite simple

Let's assume that $N_t$ people are detected at $t$, and that in $t-1$ we had $M_{t-1}$ tracked people. For each track in $t-1$, we predict its position in $t$ assuming that the displacement on the XY plane between $t-1$ and $t$ will be the same as between $t-2$ and $t-1$ (note that this prediction could also be done using other predictive filter, like Kalman, but we found it not necessary for our particular scenarios).

Then, we compute a $M \times N$ distance matrix in which the elements contain the euclidean distances between the people positions of instant $t$ and the $M_{t-1}$ predicted track positions. Using the distance matrix, objects are tracked by repeatedly:

- Find the position of the minimum of the distance matrix.
- If this distance is below a suitable threshold, the corresponding object and track are linked and the corresponding row and column removed. Notice that this maximum distance (in metres) is related to frame rate and the typical walking speed of a person.
- When no more objects can be linked, the loop ends and a new track starts for each un-matched object. All the unmatched tracks are terminated and they contain all the person information over time.

A few tracking examples can be viewed in [Oliver 2012a]. In the video, it is possible to see the label assigned to each person and the height map used for segmentation.

### 3.6.3 Tracking Evaluation

The evaluation of the segmentation and tracking module has been jointly performed using the school hall and supermarket databases presented in Appendix A, which gather simple and challenging scenarios respectively.

Since this Thesis focuses on the re-identification process, the collected databases do not contain a groundtruth with the segmentation mask of each person at each frame. For that reason, the evaluation of the segmentation algorithm can not be addressed by analysing the quality of the segmentation masks. Instead, we analyse the number of missing people and false candidates that is provided by the combination of the segmentation and tracking modules working jointly.

To that end, we define the *Precision* and *Recall* parameters as:

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

where $TP$, $FP$ and $FN$ are the true positive, false positive and false negative respectively. The $TP$ parameter identifies how many people have been correctly detected. $FN$ represents the number of people that the segmentation module has missed. On other hand, the $FP$ parameter represents the false candidates that the module wrongly estimates as people but they are not actually.

The *Precision* parameter measures the robustness of the method against false positive predictions. The higher the score is, the fewer $FP$ the method provides. On other hand, the *Recall* parameter provides information of the strength of the method against missing people. The higher the score is, the less number of missed people we have. Therefore we seek for values of 100% in both the parameters, which means that the method does not provide false positives and does not miss anybody.

Table 3.3 shows the results of the segmentation and tracking modules for the proposed databases, where the results obtained from the school hall clearly outperform those obtained for the supermarket. In the school scenario we find that $Precision = 98.85\%$, which means that only one false positive image has been generated. This false positive is shown in Figure 3.12.a. In this image, a student is raising his hand when he is near the camera. The system detects two local maxima and selects the highest one as the valid. Only one is valid because both maxima are too close together that it is not physically possible that two people are so close. Therefore, the system gets confused and considers the hand as another person, missing the real person id for those frames that the hand is raised. However, when he moves the hand down, the system is able to recover the person id and still track him with his initial id, so only a false positive has been generated, but the track is not broken in the end. In summary, no track has been broken in this scenario. Regarding the *Recall* parameter, the segmentation algorithm presents a $Recall = 100\%$ for the school database, which means that all the people have been detected and nobody has been missed.


a) b) c)

Figure 3.12: All these figures show examples of false candidates. Fore each column image, the top figure shows the id numbers of the people in the scene, whereas the bottom image shows the segmentation result. a) shows a student in the school hall raising his hand, which generates a false candidate and makes the user loose his id for several frames. b) shows a woman raising a broom and generating a false candidate but still preserving the person id. c) shows a woman raising a bagette, which causes that the person looses her id and can not recover it later.

Table 3.3: Performance of the segmentation module on the school and supermarket databases.

| | Precision | Recall | Broken Tracks |
|---|---|---|---|
| School database | 98.85% | 100% | 0% |
| Supermarket database | 97.33% | 100% | 3% |

On other hand, the algorithm performance slightly decreases on the supermarket database. The segmentation algorithm provides a good *Recall* = 100%, and the *Precission* drops to 97.33%. This decrease in the *Precision* is mainly due to the fact that people in the supermarket can raise grabbed objects that may confuse the segmentation algorithm, as depicted in Figure 3.12.b and 3.12.c where in the first case a woman is raising a broom and in the latter other woman is raising a bagette over her head. As a consequence of this, the person tracks are broken in these cases. The segmentation algorithm labels these cases as new objects, since the top of the objects are far from the head of the person that are grabbing them. When the object is returned below the height of the person head, the algorithm starts tracking the person again but with different id, which breaks down the track. However, this can not be considered as a false negative case, since the person has been tracked during more than 90% of their life in the scene. In total, in the supermarket scenario there are 5 broken tracks due to similar problems, which means that 3% of the people in this scenario have broken tracks.

Figure 3.13: Segmentation results on challenging examples from our supermarket dataset

Figure 3.13 shows a qualitative evaluation of the segmentation module for both the supermarket and school hall datasets. Figure 3.13.a shows a woman taking a shopping basket from a pile in a supermarket. It can be seen how part of the basket is included into the segmentation mask, but the algorithm is still able to disconnect it from the body. Figure 3.13.b shows a woman carrying other shopping basket. The algorithm successfully segments the person and does not consider the shopping basket. Figure 3.13.c shows a woman pushing a trolley. It can be seen how the algorithm is able to segment up to the hands, and can remove almost completely the trolley. On other hand, Figure 3.13.d shows a couple walking closely so that there is no distinction on how to segment them if using only depth information. Thanks to the height map representation, the algorithm is able to detect two maxima and can break down the connection.

## 3.7 Low-Level Person Representation

The use of the segmentation mask allows to retrieve information about the person color distribution and 3D shape for each frame of their track. For each pixel $k$ in the segmentation mask, we can extract a point $p_k$ with the following attributes:

- Position in space: $X_k^{world} = (x_k^{world}\ y_k^{world}\ z_k^{world})$
- Color: $RGB = (R_k\ G_k\ B_k)$.

The set of points $p_k$ corresponding to the $i$-th person at frame $t$ are named the *dense person pointcloud*, denoted as $\mathcal{P}_i = \{p_k\}$. Figure 3.14.c shows this dense pointcloud of a person at a particular instant.

One particular problem of using world coordinates is that they greatly change depending on the particular location of the person in the scene, which makes this coordinate system not discriminant for person re-identification. Therefore, in order to provide a better representation of the cloud that eases the person analysis and description, the 3D points are transformed from *world* coordinates to a new coordinate system centered on the person by applying a translation and rotation as follows:

$$X^{person} = R(X^{world} + T) \tag{3.5}$$

where $T$ and $R$ are the translation and rotation matrices respectively at instant $t$. Figure 3.14 shows this process, which is further detailed in Section 3.7.1.

### 3.7.1 Person-Centered Cylindrical Coordinate System

The local representation of the dense pointcloud eases the process of description of the person appearance and shape at the particular instant $t$. This local coordinate system has its origin onto the floor plane ($z^{world} = 0$) and is centered on the downwards projection of the pointcloud center-of-mass. If we name the center-of-mass of the person at instant $t$ in world coordinates as $\vec{c}^{world}$, the translation vector $T$ is set to: $T = (c_x^{world}\ c_y^{world}\ 0)$, which means that coordinate system is translated in $XY$ plane. After applying the translation to the pointcloud, data in the new coordinate system are rotated along $z$ axis, so that $x$ points towards the walking direction. Note that there is no rotation in $z$ axis, so $\hat{Z}^{world}$ still points perpendicular to floor.

The estimation of the walking direction used to angularly align the local coordinate system at instant $t$ is carried out by calculating the displacement vector of the person between $t - n$ and $t + n$ instants, where $2n$ identifies the number of frames used to calculate the direction. Note that this value may vary depending on the scenario and the camera acquisition rate used for that particular scenario. For corridor-like scenarios where people normally use straight trajectories, $n$ can be larger in order to get a smoother estimation of the person trajectory. However, for those scenarios where people move rapidly and randomly, $n$ should be smaller to rapidly update the changes in moving directions.

With this particular alignment, the front side of the person is always pointing towards $\hat{x}$ direction at all instants, independently of the trajectory. This means that people will always be identically aligned so that point coordinates belonging to a body part will always have the same position relative to the body. This person alignment is the key aspect for the future person description algorithms presented in this Thesis.

Note that cylindrical coordinates are used to represent the cloud in the person-centered coordinate system, since this coordinates eases the description of the person shape.

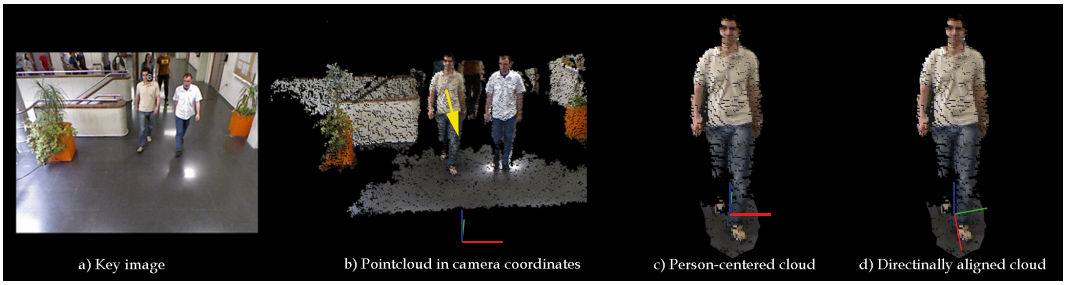



Figure 3.14: Translation and rotation of the center of coordinates according to the person center of mass and the moving direction. The $x$, $y$ and $z$ axis are represented by the red, green and blue colors respectively.

### 3.7.2 Raw Features

The person representation using a dense pointcloud provides too much information that is difficult to manage in real time. In order to make the pointcloud more handy, data is quantified in angle and height using a 3D cylindrical grid (see Figure 3.15), which is centered on the person center of mass $\vec{c}^{world}$. This quantisation consists of dividing the 3D space into a set of angular and vertical bins. For each bin, we calculate the mean color and mean radius of those points that fall into the particular bin of the grid. This representation of the data is named as the *Raw Features*, where each feature has the following attributes:

- $\overline{color}(\theta, z)$: Mean color of the pixels that fall into the corresponding bin
- $\overline{radius}(\theta, z)$: Mean radius of the pixels that fall into the corresponding bin
- $\bar{w}(\theta, z)$: Bin weight, which corresponds to the total number of pixels that voted into that bin

The resolution of the cylindrical grid can be set according to two parameters that we define as:

- $NBO$ (Number of Orientation Bins), which specifies the number of angular divisions of the quantised cylindrical space
- $NBV$ (Number of Vertical Divisions), which specifies the number of vertical divisions of the quantised space

This quantisation is used to provide a more compact representation of data following a *feature pooling* fashion [Boureau 2010], as used in many other well-known descriptors such as SIFT or SURF. This methodology compresses the information, removes irrelevant details and provides better robustness to noise and clutter. The compact representation provides a useful low-level person descriptor that can be easily handled and post-processed by higher-order descriptors that can focus on particular features.

To that end, we seek for a fine quantisation grid that provides a lighter and compact representation but still preserves the essence and distribution of the color and shape of the people. Thus, the 3D grid is build using $NBO = 64$ angular bins and $NBV = 110$ vertical bins (we assume that the maximum height for a person is $2,2$ meters). These parameters result on an angular resolution of $5,6^{\text{o}}$ and $2cm$ in height, which provide a fine resolution for describing people appearance and shape. Figure 3.15.d shows the person pointcloud represented in the quantised space using this fine grid.

One interesting characteristic of this feature extraction approach is that it is easy to automatically reconstruct the 3D volume of a person while he or she is walking in front of a camera. Since the person cloud is aligned with its trajectory, the complete shape and appearance of the person can be generated by accumulating all the 3D person appearances over time, taken from different perspectives (see

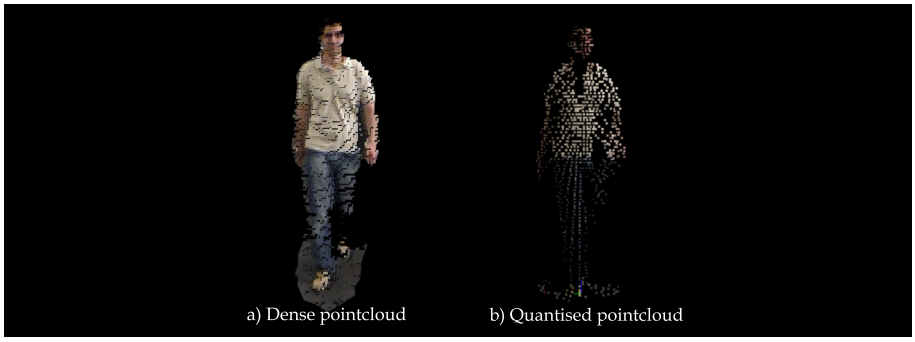



Figure 3.15: This figure shows the quantisation of a dense pointcloud using a 3D cylindrical grid. Image a) shows the dense pointcloud for a particular time. Image b) shows the quantised pointcloud into 64 angular bins.

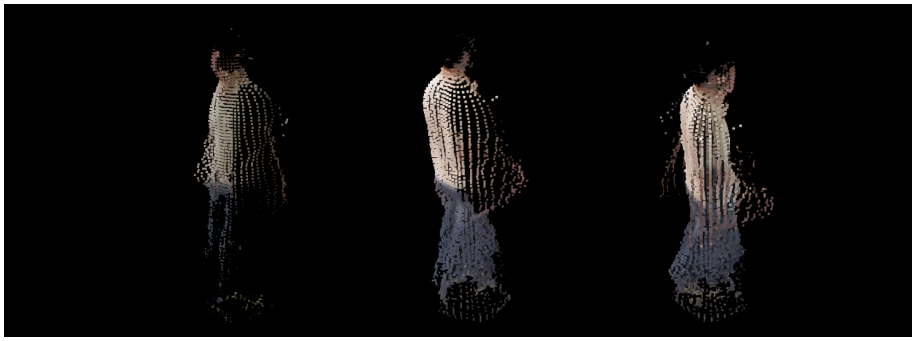


Figure 3.16: Example of volumetric person reconstruction from multiple cameras. a) shows the quantised pointcloud at a camera accumulated over time. b) shows the quantised pointcloud accumulated over time at other camera. Note that the clouds may correspond to different time periods. c) shows the merged pointclouds. It can be observed both the back and the front parts of the person fit together.

figure 3.16). It is possible to cover all the perspectives of a person given two cameras with opposed views, like in the school hall. Also, it is also possible to retrieve the 3D volume even from a single camera, when the person does a bend in the trajectory, as the example video shown on the database website [Oliver 2012b].

The raw features obtained in this pre-processing module presents three different components: time, angle and height. This raw information presents different problems. On one hand, the information describing a person over time is too heavy to be processed directly. On other hand, data presents too much correlation and noise. In this Thesis we address how to process this data to obtain a trade off between accuracy and computational time in order to make the re-identification process feasible to be embedded in commercial systems. We have studied several techniques that marginalise each of these three variables, and the results are presented in Chapter 4. Section 4.4 presents a descriptor that marginalises the temporal and angular

Table 3.4: Differences among the proposed person descriptors according to the marginalisation criteria of the input data.

| | Marginalization | | | |
|---|---|---|---|---|
| Descriptor | Time | Angle | Height | Radius |
| BodyPrint (BP) | yes | yes | no | yes |
| Cylindrical BP | yes | no | no | yes |
| Latent BP | yes | yes | no | yes |
| BoA | no | yes | no | yes |

variables, so that the descriptor of a person over time is simply represented by a single vector that contains the mean rgb values at the different heights. In Section 4.5 a descriptor that marginalises the temporal component is presented. The descriptor stores the information of the mean colors over time quantified into angular and vertical bins, so that the descriptor spatially maps the color distribution of the person. Section 4.6 marginalises the angular and radial components to consider person variations over time. Finally, Section 4.7 bases on Bodyprints and reduces the dimensionality to reduce the high correlation of the feature vector. Table 3.4 shows the classification of the proposed descriptors attending to the marginalised variable.

Chapter 4

# Person Description and Matching

## Contents

## 4.1 Introduction

The tasks of person segmentation, tracking, description and matching are the key aspects in a person re-identification process. In Chapter 3, the segmentation and tracking processes have been addressed. This chapter addresses the description and matching processes to complete the re-identification pipeline.

As it has been shown in Section 3.7.2, segmentation and tracking modules in this work provide a set of raw features that gather heterogeneous data such as color, 3D shape and temporal information of the person. These are rich features that provide complete information of the person along their tracking that definitely help in the identification process. The use of these features for re-identification purposes opens a new paradigm in this field, where 3D shape plays an important role in the process. However, these features lay on a high-dimensional space that makes them intractable because of different problems. On one hand, the problem of curse of dimensionality: when the dimensionality of the data increases, the volume of the space increases exponentially, so the available data become sparse. The analysis, searching and organisation of data in high-dimensional spaces is complex since data is sparse and dissimilar. On other hand, there are problems related to strong feature correlation, noise, outliers and missing data. Therefore, further processing is necessary to simplify the input space while preserving the most discriminative information.

The task of feature extraction in high-dimensional spaces can be handled from either a statistical or analytical point of view. From a purely statistical approach, one should set a cost function that should be minimised in order to extract the most relevant information in the training data. This process is appropriate when the number of observations is remarkably higher than the number of dimensions of the data. On other hand, the analytical treatment of the problem allows to apply prior knowledge that simplifies some of the above mentioned problems, thus reducing the amount of observations respect to a pure statistical approach. In this Thesis, since the number of dimensions of data is much higher than the number of the observed data, we decided to address the problem from an analytical point of view in a first step. Then, using the reduced feature space obtained from the analytical work, we address the problem from a statistical point.

## 4.2 Motivation and Contributions

According to the nature of the raw features extracted from the rgb-depth sensor, the input vector can be split into three major categories: color, 3D shape and time, which conform the axis of the high dimensional space. The problem under analysis is how to reduce the dimensionality. For example, it is undeniable that the time component contributes in a better characterisation of the observed person. The temporal tracking of the person provides information of all the different

appearances along time. Also, the gait analysis can be used for discriminating a person in controlled scenarios (note that in this case, the huge variability and scarceness of the data makes a major challenge to use this information for this particular end). However, temporal data might be noisy and incomplete, so it is not clear how to take the time information into consideration. On one hand, we could fuse all the temporal information into a mean vector, so that the information is averaged and outliers and noise are intrinsically removed. On other hand, we could just process each sample independently to catch all the different appearances of the person. Both approaches make sense a priori and can't be discarded without further evaluation. A similar situation occurs with the 3D shape information. It is clear that 3D shape information is useful for describing a person. However, it is important to evaluate how accurate the input information is, how many outliers are found, etc. It is also important to check the contribution and accuracy of the three spatial axis. Our prior knowledge leads us to decouple the height dimension from the other dimensions in the 3D space. All in all, a deep analysis and experimentation is required to determine which are the best features that contribute the most in the re-identification process and how they have to be processed to obtain the best re-identification rates.

The main contributions of this Thesis in this scope are the design and assessment of the following re-identification methods. These methods naturally appeared during this work as a continuous evolution towards a more accurate and robust re-identification.

1. Bodyprints: This is our first approach for person re-identification. The Bodyprints are simplified features that fully marginalise over time and partially in space, preserving height information but averaging angular and radial information. Considering a maximum person height of 2.2 meters and a step 2cm, the descriptor can be defined by a 110-component feature vector where each component represents the mean color over time for a determined height of the person. Bodyprints are simple, handy representations of people appearance that perfectly generalise the problem of description. Because of averaging in time and space, these features intrinsically solve many problems related to missing data and outliers. Part of the success of this approach relies on the high accurate estimation of the person height, which makes color information be reliably distributed in vertical bins. This fact allows the generation of distinctive vectors with spatially-distributed color features. The feature matching is carried out using different distance-based functions. In this Thesis, we used Euclidean, cosine, Mahalanobis and correlation.

2. 3D Bodyprints: These features are similar to the Bodyprints, but also use the angular information. They naturally came from the Bodyprint evaluation to mitigate some of the problems found in the first descriptor. Features rely on a discrete cylindrical space with fixed radius (imagine features laying on

a cylinder surface), where each coordinate in the space stores the mean color and mean radius over time. This spatial representation of the color allows to describe the person appearance by means of color features as a function of the angle.

3. Bag of Appearances (BoA): Given the raw features of a person, the features are just marginalised in angle and radius domains, preserving height and temporal information. This representation can be thought as having a collection of one-frame-duration Bodyprints, all put together in a bag. The advantage of this method is that it can naturally tackle the problem of temporal variability of the person appearance associated with the gait, pose, carried objects or illumination changes. In this case, using the mean appearance from different time instants would not yield any benefit because the color distribution is no longer monomodal.

4. Latent features: Bodyprints provide a general representation of the person appearance, but still fail in too much feature correlation. In order to alleviate this effect, dimensionality reduction techniques based on a probabilistic latent variable models have been applied to Bodyprints to generate the Latent Features. These are compact features generated by a probabilistic model where Bodyprint features are set as the observation. In this model, which uses the noise observation in the estimation of the latent variables, we use the temporal color variance of the Bodyprints as the noise. As a result, a reduced set of features are obtained that remove correlation and also can intrinsically cope with observation noise and missing data.

To constrain the scope of this work, we have put the focus on the person description module, which is the goal of this Thesis. The proposed methods, which require color information as input, have been generally defined to accept any color space that could be normalised against intensity variations produced by tracking in multiple camera systems. In order to provide a complete analysis of the problem, complementary evaluation regarding the selection of the color space and the color normalisation have been conducted in this work. These tests have been carried out upon the Bodyprints features, which are a general representation of the person appearance and represent the basis of all of the proposed methods. The results regarding the influence of the color space and color normalisation obtained in this study using Bodyprints have been applied to all the subsequent description methods.

Similarly, the impact of considering auxiliary information such as contextual cues or other auxiliary data have been put aside of the main discussion of this Thesis. This information, which mainly refers to time constrains in controlled spaces, may positively contribute in better re-identification rates. However, they are not directly related to the description methods and can be applied in all cases with independence of the re-identification approach. Thus, the effect of considering such information has been demonstrated in Section 4.4 using Bodyprint descriptors

and can be extended to the other approaches.

The structure of this Chapter is as follows: In sections 4.4 to 4.8, we introduce and evaluate all the proposed family of methods. Section 4.4 introduces the Bodyprint descriptor. In Section 4.5, the 3D Bodyprint descriptor has been analysed. In Section 4.6, we introduce the Bag-of-Appearances descriptor. Section 4.7 introduces the latent features, which use Bodyprints as input features. Finally, in Section 4.8 a comparison of all the proposed methods with the state-of-the-art methods is carried out.

## 4.3 Experimental Methodology

This section describes the evaluation methodology used in this Thesis to assess the performance of the proposed re-identification methods. The methods have also been compared with the state of the art works found in the literature.

In order to allow a fair comparison of methods, the evaluation strategy plays a key role. To this end, a standard evaluation methodology is presented in this section. The evaluation section is structured in two blocks: first, training and test datasets are described together with the evaluation strategy using the test set; second, the evaluation tools for presenting the re-identification rates are presented.

### 4.3.1 Training and Test Datasets

The problem of person re-identification has been typically addressed using conventional RGB cameras. However, the recent introduction of depth cameras has brought a new dimension in image processing. The pioneer research in this field has intrinsically brought the problem that there are no available public databases that can be used for the assessment of re-identification methods that use depth information. In order to be able to assess the proposed methods, we have created a public database for people re-identification using depth cameras. To our knowledge, this is the first person re-identification database that uses depth information.

The video sequences of the database were acquired in two different scenarios. One scenario is a school hall where people walk naturally along a corridor. The other scenario is a supermarket, where people behave freely with no restrictions. For both scenarios, there are two different cameras covering different areas with no overlapping field of view. A complete description about the datasets can be found in Appendix A.

A common methodology step to assess the performance in most recognition problems is to divide each dataset into two disjoint subsets:

- training set: Training sets are used for adjusting the parameters in all the proposed methods. In the training phase, we used cross-validation [Kohavi 1995].

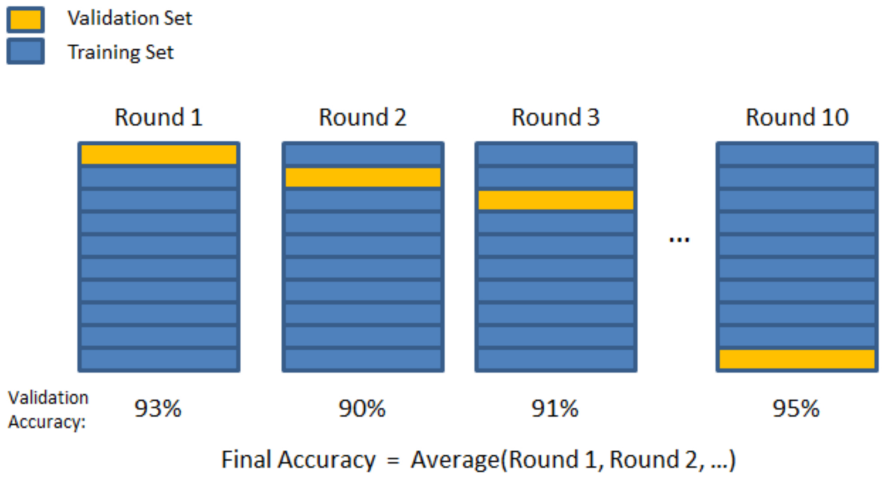



Figure 4.1: Illustration of the cross-validation process.

The objective of this technique is to define a subset to evaluate the methods in the training, in order to limit problems like overfitting and give an insight on how the model will generalize to an independent dataset. In cross-validation, the data is randomly divided into $k$ disjoint blocks of samples, usually of equal size. The training set is composed using $k-1$ blocks, and the remaining block is used as the validation set. In each block, at least one representative sample from the validation set must be included. This process is repeated $k$ times and the final result is obtained by averaging the $k$ results obtained in the iterative process in order to correct the optimistic nature of training error. Figure 4.1 shows the process of cross-validation. In this Thesis, we have used $k = 7$ blocks.

- test set: This collection of images, which do not include any sample from the training set, is used to perform the assessment of the methods using the methods parameters found in the training stage. Typically, this set is divided into gallery and probe set. A sample in the probe set is normally matched to a sample in the gallery set. In our particular case, the probe and gallery sets are independently composed by the images captured at different camera nodes. To illustrate this division, we provide an example using the supermarket dataset. In this dataset, the gallery set is formed by all the images captured by the camera at the entrance point. The probe set is formed by the images captured at the exit point.

### 4.3.2 Evaluation Metrics

In this Thesis, the standard Cumulative Matching Characteristics (CMC) and the Probability of re-identification have been used to assess the proposed re-

identification methods.

The CMC curves [Gray 2007], also known as Rank Curves, represent the probability that the correct match is in the first $r$ ranked candidates, where $r$ varies from 1 to the total number of candidates. The CMC curves are obtained in the following way: each probe image is matched to all the gallery set and an ordered ranking based on similarity is obtained for all the matches. The process is repeated for all the probe images and the average ranking is obtained. A rank $r$ indicates the number of rank candidates taken into account to search for the correct match. Rank $r = 1$ expects the correct match to be the best match in the comparison ranking.

CMC curves are of special interest in human assisted re-identification processes. In these systems, a human operator manually validates the correct match within a set of $r > 1$ ranked candidates.

In non-assisted processes, however, systems need to provide the matching rate at $r = 1$, since there is not a human operator to make the final decision. This particular case is known as the re-identification rate and shows the percentage of correct matches.

## 4.4 Bodyprints

In this section, a new appearance descriptor is presented with the name of Bodyprint. A Bodyprint can be regarded as a general representation of the person appearance over time, simplifying the input information into a compact feature vector. The general idea is that Bodyprints have the same role as fingerprints for person re-identification. They are a sufficiently distinctive set of features that allow discriminating people.

Bodyprints are the most compact representation of the raw features obtained from the rgb-depth camera (see Chapter 3). The features fairly generalyse the person appearance in time and space in a handy feature vector. The Bodyprint descriptor is chronologically the first method that we present to carry out the re-identification. Therefore, we will use the Bodyprints to address general issues like the selection of the color space, the effect of the contextual cues or to perform a deeper analysis of the similarity measurements. Note that these tests are not under the scope of this Thesis, so they will be just addressed in this section to show their general impact on the descriptors.

The following items summarise the work that will be covered in this section:

- Introduction of the Bodyprint descriptor
- Influence of the color space representation. RGB and YC spaces will be compared and discussed.
- Description and validation of a color normalisation method. The color normalisation is a key factor in all color-based description methods. We will evaluate the influence of the color normalisation in the probability of re-identification in both $RGB$ and $YC$ color spaces.
- Study the influence of different similarity measurements. We will also evaluate how the weight information, which measures the number of pixels used to extract a particular feature component, can contribute in the matching process to improve the identification rates.
- Analysis of the influence of contextual cues in the matching process. These cues are mainly related to temporal constrains that can be set in scenarios where there is an entrance and exit gates, so that one person that enters the shop necessarily leaves it from the other gate. That implies that the number of matching candidates for a query sample at entrance is greatly reduced to all those target samples that left the shop after the query sample entered the shop. The analysis will be carried out using the supermarket scenario, which gathers the most challenging sequences.

### 4.4.1 Algorithm Description

In Chapter 3 we introduced the raw features, which provide a complete description of the person appearance during a period of time. These features basically set how the color pixels are distributed over time and space 3.7.2.

The key idea of Bodyprints is that each of its components summarizes the color appearance at a different height for a tracked person. Our algorithm is similar to [Bird 2005] in the sense that it also divides a person into stripes at different heights. However, Bodyprint descriptor takes advantage of the good precision of the kinect sensor and the scene calibration that allows to precisely compute the height of all the pixels associated to a segmented person over time.

Bodyprints are used to represent the global person appearance during all the tracking. For that reason, the averaging of all pixels over time for a particular height bin is important because it provides generalisation, removes missing data and partially removes noise easily.

The height axis $z$ is discretized at steps of 2cm. We assume a maximum person height of 2.2 meters, producing $N_h = 110$ vertical bins in the Bodyprint descriptor. For a person $i$, the Bodyprint is defined as:

$$F_i(z_h) = \left[\overline{color}_i(z_h), w_i(z_h)\right] \tag{4.1}$$

where $\overline{color}_i(z_h)$ is a column vector with information of the color, calculated marginalising the variables $t$, $\theta$ and $r$ from the raw features, $color_i(t, z, \theta, r)$; and $w_i(z_h)$ is a weight vector that measures the number of pixels contributing to calculate the mean color for a particular height, as follows:

$$\overline{color}_i(z_h) = \frac{1}{w_i(z_h)} \sum_t \sum_{\theta,r} \overline{color}_i(t, z, \theta, r)\, w_i(t, z, \theta, r) \tag{4.2}$$

$$w_i(z_h) = \sum_t \sum_{\theta,r} w_i(t, z, \theta, r), \quad where \quad 0 \leq h < N_h \tag{4.3}$$

Basically, each bin of the color vector $\overline{color}_i(z_h)$ stores all the pixels that spatially and temporarily belong to a particular height. The count vector $w_i(z_h)$ is a measure of the reliability of the values in the Bodyprint. This gives more importance to those body parts that are more visible to the camera, and therefore are more reliable. It is possible that some parts of the body have low number of pixels due to problems derived to the quality of the depth sensor or the camera perspective.

Figure 4.2 shows the process of Bodyprint extraction from temporal, angular and radial marginalisation.

The first column represents a key frame in the sequence containing a person.

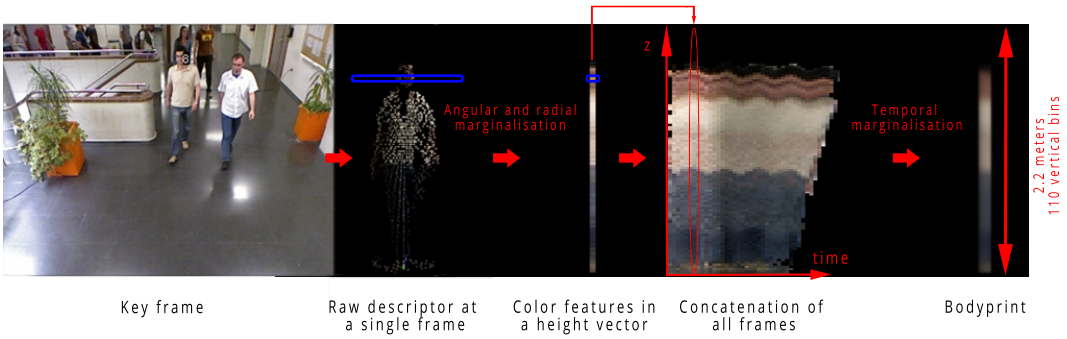



Figure 4.2: Process of Bodyprint generation. The first image shows a key image during the person tracking. The second image shows the pointcloud at a single frame. The cloud is marginalised in angle and radius to generate the third image, which represents a column vector storing color information. Image four shows the concatenation of all the column vectors over time. This representation is named Temporal Signature of the person. When the Temporal Signature is marginalised in time, it generates the Bodyprint feature vector, shown in the last image.

The second column shows the raw descriptor at a single frame. The $z$ axis is split into 110 vertical bins. For each bin, the mean color is calculated by summing all the pixels in that volumetric slice, marginalising against angle and radius. The process is repeated for each bin to create the color feature height vector, as shown in the third column.

The fourth column shows an intermediate state of features, named temporal signature of the person, $\overline{color}_i(t, z_h)$. Although this concept will be further discussed in Section 4.6, we will briefly introduce it in this section. The temporal signature is calculated by appending the height vectors from all the frames of the tracking. In this representation, the horizontal axis represents time and the vertical axis represents the height. Black values indicate that no pixels were found at that particular height and instant. Note that in these examples people are walking downwards and for this reason the feet are the first part that leave the image (right portion of temporal signature). Notice also that the width of the signatures varies because it depends on the time that a particular person was visible (which in turn may depend on the person walking speed). The temporal signatures exhibit a small ripple caused by variations of height while walking. This ripple could be compensated, however our experiments show that the impact of this artifact in the final performance is almost negligible.

The fifth column shows a Bodyprint $\overline{color}_i(z_h)$, obtained by marginalising in time, angle and radius.

Figure 4.3 shows more examples of Bodyprint features for five different people. For simplicity, only the key image, the temporal signature and the Bodyprint are depicted. It is easy to see how distinctive a Bodyprint is. The mean color distribution

along person height provides a fair representation of a person over time. Besides, the person height as well as the trousers and shirts height can be easily distinguished from the descriptor. The high precision in the estimation of the person height is the strength of this method in comparison to the one from Bird et al. [Bird 2005].

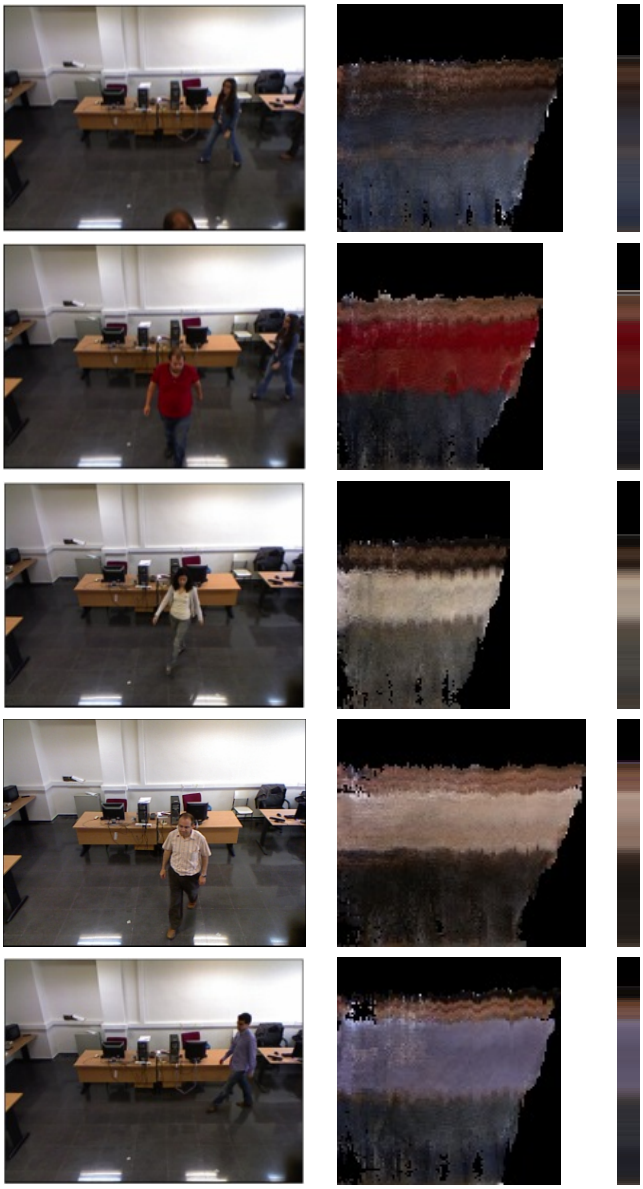


Figure 4.3: The first column shows a key frame of each person during their tracking. The second column shows the temporal signature of each person. The third column shows the Bodyprints.

Just to illustrate what happens in the case of a temporary occlusion, in Figure 4.4 we can see a temporal signature containing a brief occlusion: the woman wearing a white shirt is totally occluded by the man wearing a blue shirt during a second.

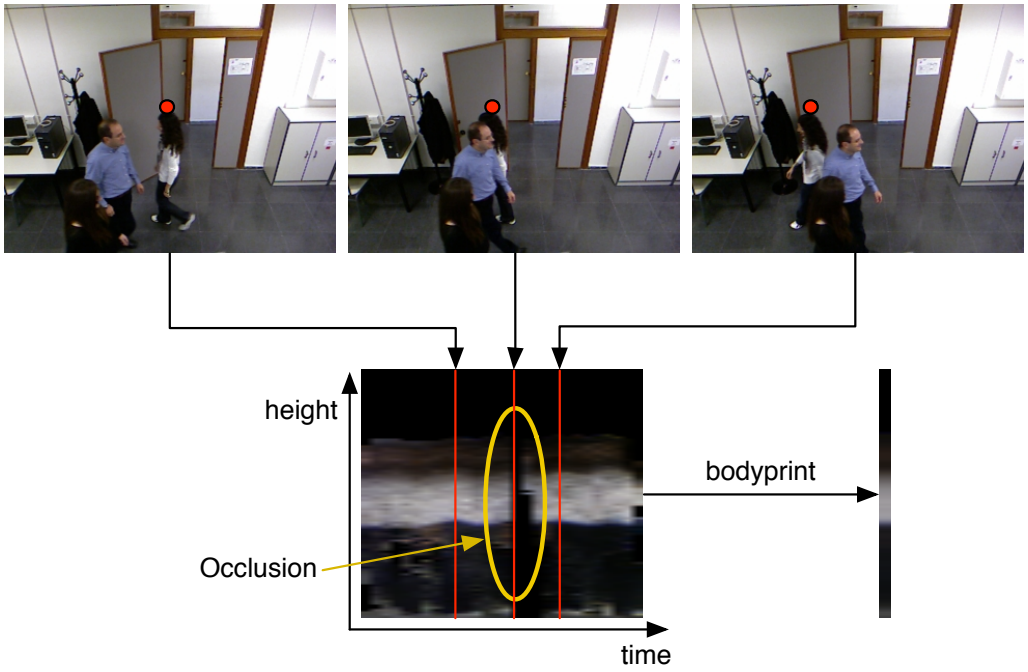



Figure 4.4: Example of track with temporary occlusion. Above: frames at different times. Below: temporal signature and Bodyprint.

Notice the gap in the temporal signature during the occlusion. However, thanks to the tracking the person description is not divided in time. It is possible to build the whole Bodyprint merging the information of several instants, removing the gap with missing data. Figure 4.4 also shows a partial occlusion that is propagated during all the person tracking. The man with the blue shirt is partially occluded all the time by the woman dressing in black, but since we have information for every height, the Bodyprint is fully formed. On the same scene, the lady closer to the camera is only visible in her upper part. This would yield a Bodyprint with some missing portions (the lower part). Matching Bodyprints containing missing portions is possible as it will be explained in next section.

All in all, Bodyprint features are distinctive features that can perfectly gather the general appearance of a person over time and intrinsically solve problems associated with temporal discontinuities, missing data or high feature correlation.

### 4.4.2 Color Representation and Normalisation

The use of color cues for person re-identification provides distinctive and relevant information about the person appearance [Kviatkovsky 2012].

Working with color information is not easy, though. Colors captured by cameras suffer from unknown, variable illumination conditions that may affect the scene globally and locally. Therefore, both the selection of an appropriate color space and

the use of effective normalisation techniques are mandatory tasks to obtain robust color-based features.

In the following subsections, a detailed description of the used color spaces and a proposal of a robust color normalisation technique is presented.

#### 4.4.2.1 Color Models

Color models are numerical ways of describing colors in a particular coordinate system. Any color defined using a color model is represented as a single point in the subspace that model defines. In these spaces, each color pixel value can be algebraically transformed to a different color space with different properties while keeping the original image structure. There are different color spaces depending on the type of the transformation. Typically, these spaces are represented by three channels: $C_1$, $C_2$ and $C_3$. Depending on the color space, the geometrical distribution of the color samples in an object may vary. Therefore, depending on the nature of the input data, different color spaces may provide better representation of the data. Some conventional color spaces used in color-based person re-identification are: $RGB$, $YC_BC_R$, $HSV$, *Opponent color...* Despite the effort in color analysis for person re-identificaton, the search of the optimal color subspace is still an open problem.

In this Thesis, we propose the use of $RGB$ and $YC_BC_R$ color spaces for person re-identification, together with some variations on these color spaces. Using the $RGB$ color model is convenient because the human visual system works in a similar way, it is easy to work with it and it is the most extended color space. The $YC_BC_R$ space is also a very used color space in the image processing. It decouples the information of intensity and chroma.

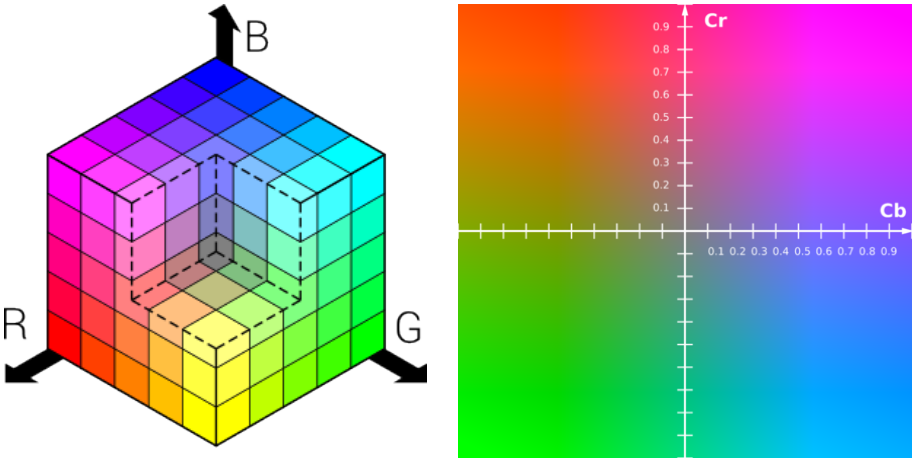



Figure 4.5: Left image: $RGB$ color model. Right image: $C_BC_R$ plane detail at a constant luma Y'=0.5 for $YC_BC_R$ color model

#### $RGB$ Color Model

The $RGB$ color model, which is commonly represented by a cube (Figure 4.5), is an additive model that uses Red, Green and Blue primary colors. It is the most used color representation scheme in Computer Vision and Computer Graphics because the human visual system works in an similar way.

Although this model does not present remarkable discriminative properties for person description problems, it is considered as a fundamental model in the image processing field, since many different color spaces can be algebraically derived from it.

#### $YC_BC_R$ Color Space

The $YC$ family of color spaces are color representation systems derived from $RGB$ primaries that separately use the luminance and chrominance to encode the color information. They describe the color taking into account the human perception. This family of spaces is widely used in image and video systems to reduce transmission bandwidth. The human eye is more sensitive to brightness information, so the luminance channel has significant more impact on the perceived image than the other chrominance channels. That allows to compress separately luminance, allowing higher compression in the chrominance.

The $YC_BC_R$ is created from the corresponding gamma-adjusted $RGB$ source using two defined constants $K_B$ and $K_R$ as follows:

$$\begin{aligned} Y' &= K_R * R' + (1 - K_R - K_B) * G' + K_B * B' \\ C_B &= \frac{1}{2} * \frac{B' - Y'}{1 - K_B} \\ C_R &= \frac{1}{2} * \frac{R' - Y'}{1 - K_R} \end{aligned} \tag{4.4}$$

where $K_B$ and $K_R$ are ordinarily derived from the definition of the corresponding $RGB$ space, and $R'$, $G'$, $B'$ nominally range from 0 to 1. In this Thesis we followed *ITU-R BT.601* conversion, where $K_B = 0.114$ and $K_R = 0.299$.

### 4.4.2.2 Color Normalisation

In order to obtain robust features that are invariant to local and global illumination changes, a previous step of color normalisation is required. Many authors obtain different mean values for each color component and compensate each channel independently. However, we think that using a single mean value is useful to distinguish among colors (green vs. red for instance).

The proposed normalisation algorithm is computed using a weighted mean from $R$, $G$ and $B$ color channels. The weights are given by the variable $w_i(z)$ 4.3 and represent the amount of pixels that contributed in generating that color when marginalysing the raw features in time and space. The normalisation factor $M_i$ is the mean bright value for person $i$ over time and space, and is calculated as:

$$M_i = \left( \sum_z w(z) \frac{\overline{R}_i(z) + \overline{G}_i(z) + \overline{B}_i(z)}{3} \right) / \left( \sum_z w(z) \right) \tag{4.5}$$

where $\overline{R}_i(z)$, $\overline{G}_i(z)$ and $\overline{B}_i(z)$ are each of the color components of $\overline{RGB}_i(z)$.

$M_i$ is a constant value that is removed to all color values in the feature vectors. Therefore, according to the general equation defined in 4.3, the normalised color features particularised for $RGB$ and $YC$ color spaces are:

- $\overline{RGB}'_i(z_h) = \overline{color}_i(z_h) - M_i = \overline{RGB}_i(z_h) - M_i$
- $\overline{YC_bC_r}'_i(z_h) = \overline{color}_i(z_h) - M_i = \overline{Y}_i(z_h) - M_i$

Note that in the case of $RGB$ normalisation, all the channels are subtracted the same value $M_i$. This allows to preserve the relative value distribution among channels. If we normalised each channel independently with different mean intensities, we would loose this distribution.

### 4.4.3 Evaluation

In this section, we define, analyse and assess some general conditions that will be used in the Thesis by other methods. We evaluate different color spaces that can be used by any of the person description methods. Also, the proposed color normalisation approach will be assessed, together with the different matching metrics available for feature comparison. The effect of considering contextual cues in the re-identification process will be addressed as well.

The evaluation has been carried out using the two databases described in Appendix A. The training set has been used for parameter configuration. The test set has been used to evaluate the algorithm performance. The Probability of re-Identification has been used in the algorithm validation.

#### 4.4.3.1 Color Space and Matching Metric Analysis

This section analyses the influence of the $RGB$ and $YC_bC_r$ color spaces in the probability of re-identification in different scenarios. Also, the importance on normalising the features is demonstrated using the proposed normalisation method. Tables 4.1 and 4.2 show the comparison of the proposed color spaces against the different matching metrics proposed in B.

The columns represent the different formats of the color spaces. The variable $\overline{RGB}$ indicates that the color was processed in $RGB$ space and has not any sort of normalisation. The variable $\overline{RGB}'$ indicates that the color was processed in $RGB$ space and it was normalised by subtracting the mean intensity as it is shown in equation 4.4.2.2. The variable $\overline{YC_bC_r}$ indicates that the color was processed in $YC$ space and has not any sort of normalisation. The variable $\overline{YC_bC_r}'$ indicates that the color was processed in $YC$ space and it was normalised by subtracting the mean intensity as it is shown in equation 4.4.2.2.

Table 4.1: Analysis of the color space for different matching techniques in the School Hall scenario.

| Matching metric | $\overline{RGB}$ | $\overline{YC_bC_r}$ | $\overline{RGB}'$ | $\overline{YC_bC_r}'$ |
|---|---|---|---|---|
| Euclidean | 60.2 | 59.8 | 84.3 | 86.1 |
| Weighed euclidean | 67.6 | 65.9 | **92.4** | **94.3** |
| Correlation | 61.5 | 60.2 | 86.1 | 86.1 |
| Weighed correlation | 65.9 | 67.6 | **94.3** | **94.3** |
| Cosine | 59.3 | 52.6 | 82.9 | 84.0 |
| Weighed cosine | 64.8 | 63.3 | **94.3** | **92.4** |
| Mahalanobis | 65.0 | 67.6 | 90.8 | 89.9 |

Table 4.2: Analysis of the color space for different matching techniques in the Supermarket scenario.

| Matching metric | $\overline{RGB}$ | $\overline{YC_bC_r}$ | $\overline{RGB}'$ | $\overline{YC_bC_r}'$ |
|---|---|---|---|---|
| Euclidean | 18.4 | 17.9 | 41.5 | 40.7 |
| Weighed euclidean | 22.1 | 21.2 | **52.6** | **51.3** |
| Correlation | 15.9 | 15.9 | 39.8 | 40.7 |
| Weighed correlation | 21.5 | 18.6 | **50.7** | **52.6** |
| Cosine | 16.4 | 16.4 | 42.4 | 45.0 |
| Weighed cosine | 19.2 | 20.7 | **51.3** | **51.9** |
| Mahalanobis | 16.4 | 15.6 | 48.8 | 49.6 |

The rows represent the following different similarity measures: correlation, euclidean distance, mahalanobis distance and cosine distance. For all of them, the weighted version of the matching metric has been considered, as described in B, where the weights are given by the variable $w_k(z)$ 4.3.

In general, the results show that the re-identification rates in the school hall are higher than in the supermarket. The reason is that the first one is a simple scenario that gathers people walking straight along a wide corridor. Cameras are placed with similar perspective and the lightning is very similar. Contrary, the supermarket scenario gathers more complex situations in a real supermarket, where people do not necessary follow linear trajectories and may carry objects.

A more detailed analysis of the data confirms that the features normalisation is essential for both color spaces to get the best identification rates. Note that the terms $\overline{RGB}'$ and $\overline{YC_bC_r}'$ denote the normalised features. The explanation is simple: the normalised color features are more robust to global and local illumination changes. Global changes are related to fluctuations of the mean intensity over time due to changes in the perspective of the light source in relation to the body and the camera. These changes may occur during the tracking at one single camera (Figure 4.6 top) or also at different cameras (Figure 4.6 middle and bottom). Local changes

are related to variations of a particular color at different heights of the person, for the same instant (Figure 4.6 middle).

In all the cases, taking into account the weight information $w_k(z)$ positively contributes in the re-identification rates. The weight information measures how reliable the color information at each descriptor component is. Therefore, those color features that have been calculated from a fewer set of points are given less importance in the matching. This makes the Bodyprint robust against outliers, as shown in Figure 4.7, where we can see an example where the weight helps. The image shows a woman at the exit row of the supermarket raising a loaf of bread, which is over passing the head. That creates some noise in the Bodyprint. However, the associated weights to those feature components let us know that this information is not reliable, so it will not given importance in the matching.

On other hand, we did not find an optimal color space for the re-identification. Both normalised color spaces yield similar results. Similarly happens to the matching metrics. The weighted euclidean and weighed correlation provide the best results. Since the purpose of this Thesis is not to demonstrate a closed solution of person descriptors but to compare the different proposed descriptor methods, the selection of the optimal color space and matching metric is not critical. Anybody willing to use the descriptors can use any color space and matching metric depending on their needs and scenarios. For simplicity and ease of understanding, from here on we decide to work with the $\overline{RGB}$ color features.

Focusing the analysis in the $RGB$ color space, we show some results of correct and incorrect matches. Figure 4.8 shows good matches. The first two rows show results from the school hall; the next two rows, from the supermarket. Note that there are matches that are difficult even for the human eye.

Figure 4.9 shows wrong matches. The first rows belongs to the school hall. The next two rows, to the supermarket. In the school hall, the matching fails because the backback of the man in the right, which altered the person appearance in the front and back views, so the person viewed from its back is more similar to other person in the database.

The next two rows show wrong matches at the supermarket. The first of these mismatches shows a man dressed in gray color that is mismatched at the exit with other man with similar colors and height. The second row shows two people dressing in black colors.

#### 4.4.3.2 Contributions to the Algorithm

In this section we analyse some cues that may contribute in the matching.

On the one hand, the person height itself is a robust biometric feature when correctly measured. In non-intrusive scenarios, accurately measuring the person height is a challenging task. Bodyprints are calculated by taking the mean person

appearance over time, so the resultant vector has the mean person height. Since this information may be relevant, we can directly consider the height difference into the matching function using an independent factor, so that the similarity $\mathcal{S}(i,j)$ between these two people is defined as follows:

$$\mathcal{S}(i,j) = \mathcal{M}(F_i, F_j) e^{-\frac{\Delta H}{\nu}} \tag{4.6}$$

being $F_i$ $F_j$ the Bodyprints of person $i$ and $j$ respectively for any of the proposed color spaces. The first term measures the appearance similarity using any of the proposed similarity measures and the second term penalizes the difference in height, $\Delta H$, between the two people. The constant $\nu$ controls how the confidence decreases and its value was empirically determined ($\nu = 4$cm). Note that in the case of correlation and cosine measures, the similarity values are mapped to a positive range $[0, 1]$ so that the penalisation factor can be naturally applied according to the equation 4.6.

On other hand, applying temporal constrains when possible may also improve the identification rates since the target dataset may be significantly reduced. In the case of the supermarket, we can make the assumption that the maximum time that a person stays in the shop is 1h. With that restriction, we can directly remove those people in the database whose stay period exceeds that amount of time. In the database of the supermarket, where the recordings cover a period of two days, this information is really useful. However, in the school scenario this temporal constrain can't be applied. Note that it was a controlled scenario and the recording time was short.

In general, the results confirm that the matching rates improve when considering the height information and temporal cue. In the case of the school, the improvement is significant since the error is halved when using the height information. In the supermarket, though, the use of the height information does not provide too much benefit. The reason is that in this scenario it is more difficult to extract the actual person height due to the nature of the scenario, where people can bend to take objects, for example. Regarding the use of the temporal constrains in the supermarket, there is an improvement of 6% in the re-identification probability. Although this improvement is noticeable, the impact of considering the temporal cues is not that relevant if we think that the number of possible matches per query person has been drastically reduced when applying the temporal constrain. This fact reveals that the matching is mainly influenced by the properties of the descriptor itself. Besides, evaluating the probability of re-identification applying temporal restrictions does not provide a real estimation of the quality of the description method itself. Therefore, although the use of this information improves the re-identification rates and is strongly indicated in commercial systems, this information will not longer be used in this Thesis to evaluate the quality of the proposed methods.

Table 4.3: Contribution of the person height and temporal information cues in the Probability of re-Identification. The first column identifies the scenario. The second column shows the matching rate with no other cues. The third column shows the influence of considering the differene of the people heights as a factor in the matching function. The forth column shows the effect of applying time constrains. The fifth column shows the combination of both cues. In all cases we are using RGB color space and Euclidean distance. Please note that in the school it is not possible to use time constrain.

| Scenario | $\overline{RGB}'$ | $\overline{RGB}'$ + h | $\overline{RGB}'$ + t | $\overline{RGB}'$ + h + t |
|---|---|---|---|---|
| School hall | 92.4% | 96.2% | n.a. | n.a. |
| Supermarket | 52.6% | 53.05% | 58.11% | 58.11% |

### 4.4.4 Conclusions

In this section, the novel Bodyprint feature descriptor has been presented. The method provides a good generalisation of the person appearance during time into a compact vertical feature vector by removing the temporal, angular and radial dependence. This vector faithfully represents the average person appearance over time and intrinsically removes noise.

In the assessment, we have first evaluated some general aspects, such as the influence of the color space, the importance of the weighting vector, the matching metric or the influence of the contextual cues. Note that some of these points are not under the scope of this Thesis, but are interesting to be addressed since they are common to all the family methods presented in this work.

Two color spaces $RGB$ and $YC_bC_r$ have been evaluated. The results confirm the importance of using a color normalisation to remove local and global illumination changes. Both color spaces provide similar and acceptable results when used with the normalisation. Regarding the similarity measure, four different metrics have been presented, together with their complementary weighted versions. The weight vector provides information of how reliable a color feature is. Note that this vector directly measures the number of pixels that contributed to the averaged color. The results show that the weighted euclidean and the weighed correlation similarity measures provide the best performances, demonstrating that the weight vector plays a key role in the matching process. Note that anybody willing to use the Bodyprint descriptor can select a different color space and matching metric according to their needs or preferences.

On other hand, we have evaluated the influence of the contextual cues, such as taking benefit from time constrains. The results demonstrate that using temporal information significantly improves the re-identification score, since it directly reduces the number of matching candidates. This information will not be used in this Thesis, though its use is recommended in real applications to increase the scores.

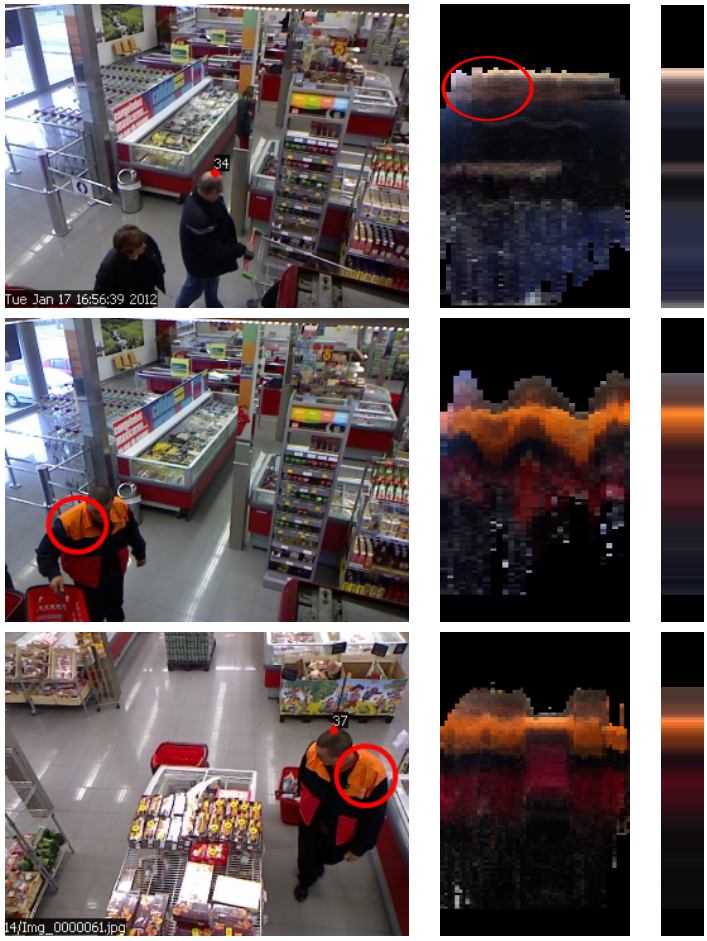



Figure 4.6: Examples of illumination changes. The first column shows a key image of the person tracking. The second column shows the temporal signature. The third column shows the Bodyprint, without illumination compensation. Top row shows an example of how the illumination changes during the tracking. In the temporal signature, the color at the head and the shoulder varies during time. The middle row shows local illumination changes. For a particular instant represented in the key frame, the orange color on the shoulders is not homogeneous, although it should. The bottom row shows the person of the middle row at the exit to illustrate how the person Bodyprint at extracted by different cameras may vary.


Figure 4.7: Outlier removal example. The first row shows a key frame of the person during the tracking. In the second row, from left to right: the temporal signature; the weights over time; the Bodyprint; the accumulated weight vector. The image shows a woman at the exit row of the supermarket raising a loaf of bread. The bread is segmented as part of the person. However, the accumulated weight factor indicates that its contribution in the Bodyprint is irrelevant.

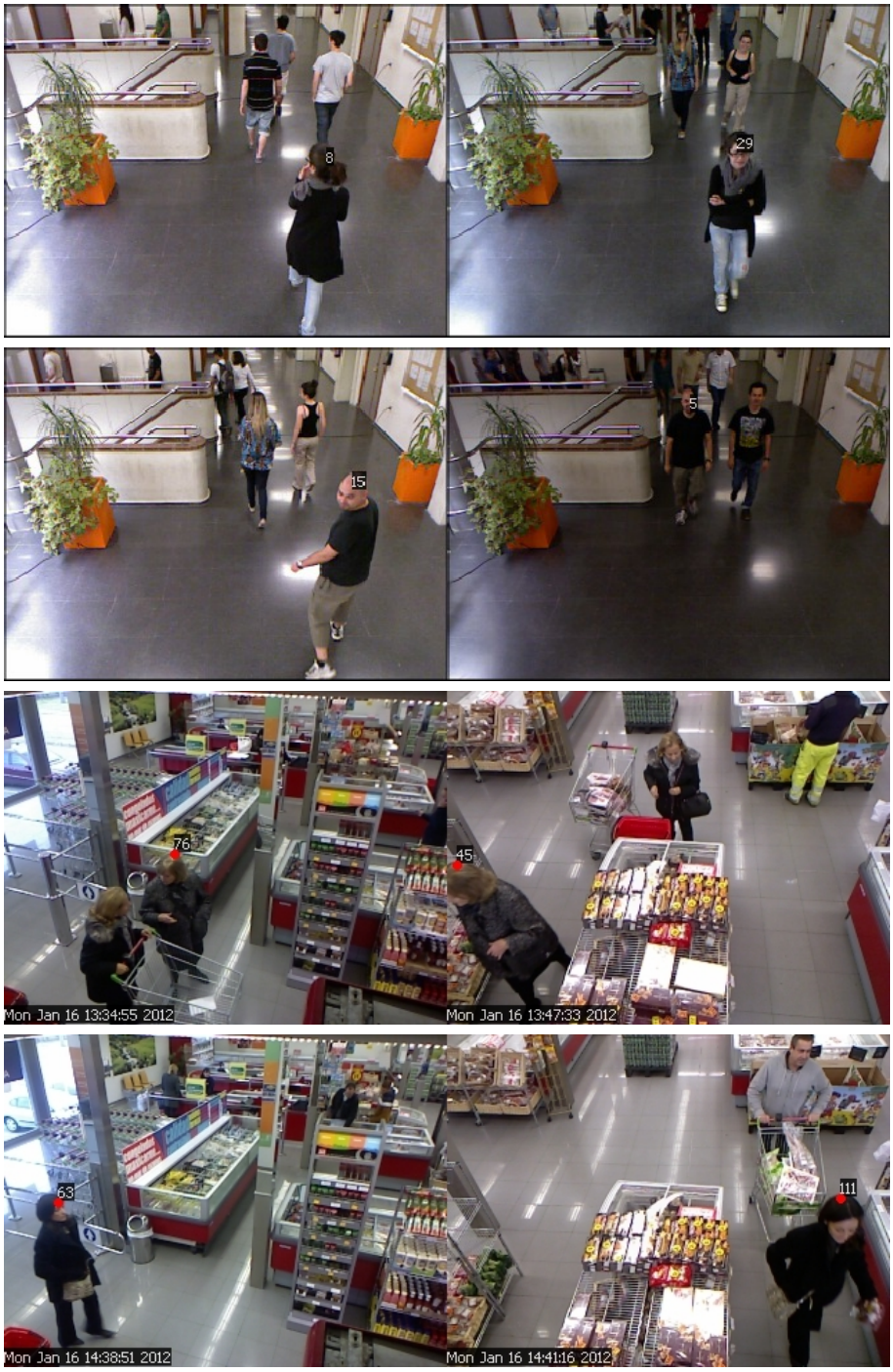



Figure 4.8: Examples of good matches using normalised *RGB* color features and Euclidean distance for matching. The first two rows show examples of matches at the school hall. The next two rows, at the supermarket.

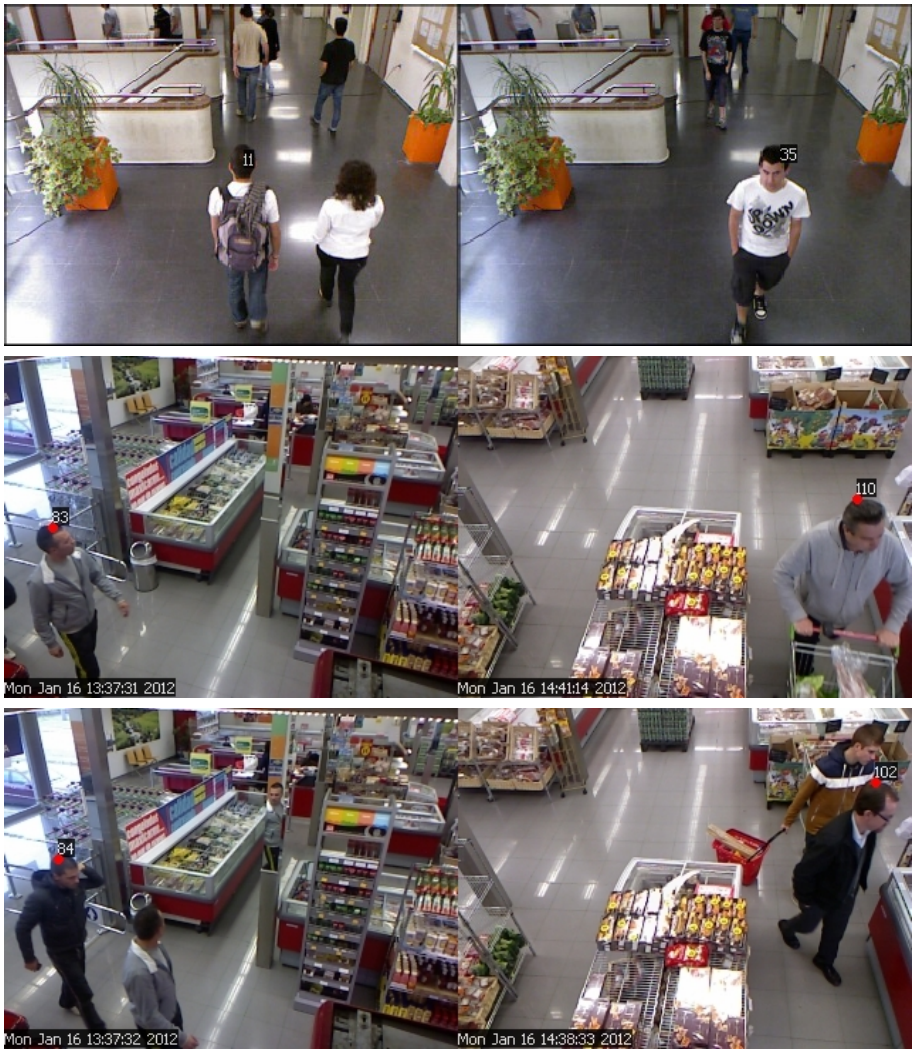



Figure 4.9: Examples of wrong matches. The first row shows an example of wrong match at the school hall. The next two rows, at the supermarket.

## 4.5 3D Bodyprints

In the previous section, we presented the Bodyprint descriptor, which we defined as color feature vectors that distribute the overall color information of a person along the person height. The descriptor has been demonstrated to yield good results in terms of re-identification probability. However, the main lack of the Bodyprint descriptor is that it fails in capturing the angular distribution of the color features.

In this section, we introduce the 3D Bodyprints, an enhancement to the Bodyprint descriptor that uses a cylindrical grid to store color features. An interesting capability of this new descriptor is that it captures the angular distribution of the features, which allows fair comparison of people at cameras with different perspectives by just considering the common body parts seen by the cameras.

The following items summarise the points that will be covered in this section:

- Introduction of the 3D Bodyprint
- Parameter analysis. Study of the relevant hyperparameters related to the 3D bodyprint
- Algorithm assessment

### 4.5.1 Algorithm Description

In Section 3.7 we introduced the raw features for person re-identification, which were denoted by $\mathcal{P}_i(t, \rho, \theta, z)$, being $i$ the $i$-th person in the database. The raw features can be represented as a temporal accumulation of dense 3D point clouds where each point encodes color information from the person appearance. These features are represented in cylindrical coordinates using 64 angular bins, each bin containing the color information in RGB format. As stated in Section 3.7.2, these features need to be yet simplified in order to make the data more handy and reduce the feature correlation and noise.

The Bodyprint features presented in Section 4.4 greatly simplify the raw feature vector by removing the dependency with time, angle and radius, just storing the mean color distributed in a vertical axis.

The 3D Bodyprint descriptor follows a similar strategy to the Bodyprint. Features $\mathcal{P}_i$ are also packed into a compact vector $RGB_i^{3D}(\theta_a, z_h)$ that has removed the dependency with time and radius variables, but preserves the angular information:

$$RGB_i(\theta_a, z_h) = \sum_t \sum_\rho \mathcal{P}_i(t, \rho, \theta, z) \tag{4.7}$$

To marginalise the radius component, for a particular time instant, the weighted mean color of all the bins laying in the same angle is calculated for each particular

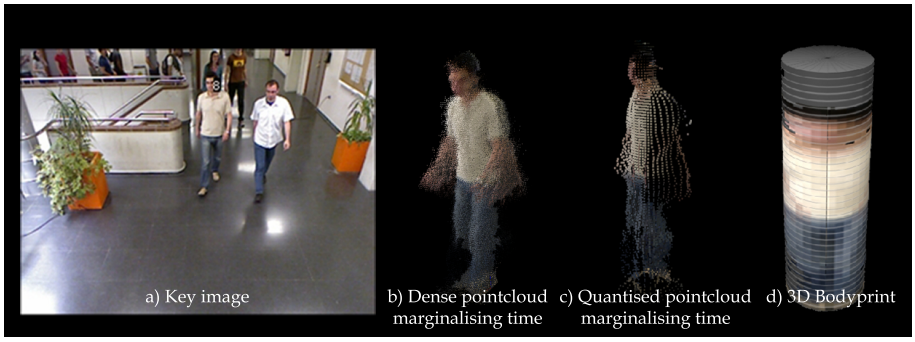



Figure 4.10: 3D Bodyprint composition. Image a) represents the a key image of the target person during the tracking; image b) represents the dense pointcloud accumulated over time; image c) shows the raw features using 32 angular bins and 110 vertical bins; image d) shows the 3D Bodyprint, obtained from image c) after removing the radial dependence.

height. On the other side, the marginalisation in time is performed as in the case of the Bodyprints: for each vertical bin, we calculate the weighted mean color over time. The result is a feature vector that can be considered as a cylinder grid, where the features are mapped on the surface (see fig 4.10-d).

As in the case of Bodyprints, the weight matrix $w(t, \rho, \theta, z)$, which stores the number of image pixels associated to each spatial bin of the descriptor, plays an important role. The marginal weight can be generated as:

$$w_i(\theta_a, z_h) = \sum_t \sum_r w_i(t, z, \theta, r) \tag{4.8}$$

Finally, in order to deal with the angular information, we introduce the term $NBO$, defined as the Number of Orientation Bins, which determines the number of angular sections in which the cylinder is divided.

In this way, the 3D Bodyprint descriptor of $i$-th person can be defined as:

$$\begin{aligned} F_i^{3D}(\theta_a, z_h) &= \left[RGB_i(\theta_a, z_h), w_i(\theta_a, z_h)\right], \\ where \quad & 0 \leq a < NBO, \quad 0 \leq h < N_h \end{aligned} \tag{4.9}$$

The figure 4.12 right shows the descriptors of three different people in two different scenarios. The first column shows the key image. The second column shows the raw descriptor marginalised in time. The third column shows the 3D Bodyprint, extracted after marginalising in the radial axis. The forth and fifth columns are a 2D representation of the 3D features. Respectively, these columns depict the weight and color distribution in height and angle, where the vertical axis represents the height and the horizontal axis the angle. The angular bin $\theta = 0(0^o)$ corresponds to the left part of the horizontal axis of the images. The angular bin $\theta = 32(360^o)$

corresponds to the right part. Note that the faces of the individuals normally appear in these angles because the orientation of the descriptor at each frame is normalised according to the moving direction, which normally coincides with the face orientation. Even in the case of the man with the red jacket, the face appears in the correct position. Note that in this case the man changes his orientation while he is being captured by the system. Therefore, this man presents a richer descriptor that has information in almost all the angular bins.

Depending on the scenario, there might be an optimal $NBO$ value that makes the person appearance more discriminant against the others, maximising the re-identification rate. The analysis of the optimal value of this hyperparameter needs to be carefully tuned up. In order to reduce the number of angular bins, the raw features are downsampled using weighed bilinear interpolation to preserve reliable information at any angular bin.

The $NBO$ value may range from 1 to 64. Notice that $NBO = 1$ directly represents the Bodyprint descriptor presented in Section 4.4. $NBO = 2$ corresponds to a simple representation that just gathers the front and rear appearance of a person. $NBO$ values higher than 16 bins are not necessary because finer angles would be too noisy for several reasons:

- The spatial and color input information is not too accurate due to the sensor precision. The commercial sensor used for these experiments (Kinect 1) provided 640x480p RGB images and 320x240p depth images. Given the position of the cameras, which were placed to widely cover the entrance and exit paths, the size of a person in a rgb image was about 100x220 pixels, halved in the depth image. With that resolution, it is difficult to extract a detailed description of the person appearance. Moreover, the depth sensor accuracy at 5 meters distance is about 7cm, as stated in [Khoshelham 2012]. This low accuracy directly affects the spatial distribution of the person point cloud using the cylindrical representation, as discussed in the following items.

- The orientation of the pointcloud at each frame is performed based on the estimation of the person trajectory. This estimation might not be accurate enough for several reasons. First, the accuracy of the depth sensor at 5 meters distance is really low, so an error about 7cm may drastically change the orientation of the point cloud at each frame. Second, the trajectory is calculated using a low-pass filter to smooth the trajectory, which may not fully cope with small variations in the person trajectory.

- The center of gravity of the person point cloud may slightly vary depending on several reasons. As explained in Section 3.7.2, the center of coordinates for representing the person point cloud is set according to the center of mass of the cloud, and it is placed onto the floor. Since we may not have the complete spatial representation of the person, the estimation of the center of mass is approximate. In order to compensate the small bias, we empirically set different offsets depending on the particular perspective of each camera.

Therefore, depending on the portion of the person that is seen by the camera, a bias in the estimation of the center of gravity may occur, producing that the features are mapped to wrong angular bins. This problem increases when the number angular bins increases.

Figure 4.11 shows the person appearance for different $NBO$. It can be appreciated that the general person appearance is always maintained for all the different $NBO$, although the texture details for those lower $NBO$ values have been lost.

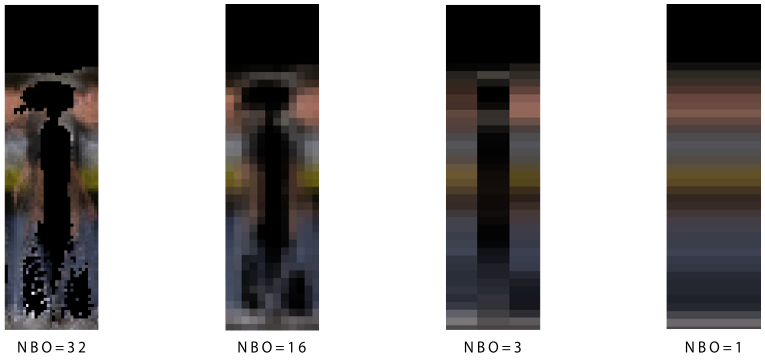



Figure 4.11: Detail of the unrolled color features of a particular 3D Bodyprint varying NBO.

### 4.5.2 Evaluation

The proposed 3D descriptor is a rich feature vector that captures the person appearance in space. However, although the dependence with the temporal and radial domains has been removed, yet it is a high-dimensional feature vector that is difficult to handle. Therefore, this experimentation targets in finding the optimal $NBO$ for the given scenarios. In these tests the $NBO$ values will range from 1 to 16 bins.

Figure 4.13 shows the results of the experimentation. The vertical axis shows the Probability of re-identification. The horizontal axis shows the different $NBO$ values. Both school hall and supermarket training databases have been used for the evaluation.

On one hand, the Probability of re-identification in the school rapidly gets the 100% of score for $NBO = 2$ and it is quite stable for higher number of angular bins. On other hand, the Probability of re-identification in the supermarket does not improve with higher $NBO$ values.

From the first case, we can see that increasing the number of bins helps in the re-identification. When the person trajectory is stable and the input data is accurate, augmenting the angular bins helps in the identification. On other hand, the tests in the supermarket show that increasing NBO does not help. The explanation of the behavior in the latter case may be directly related to the alignment phase of the cylinder, which depends on the estimated trajectory and the

estimated center of mass. A small variation in this orientation produces a shift in the alignment of the angular descriptor. Remember that for each frame, the point cloud needs to be aligned according to the moving direction in order to be able to average the color at each bin over time. Moreover, people in the supermarket interact with other objects in the scene (the trolley, the shopping basket...), which directly affects to the estimation of the center of mass of the point cloud. A small variation in that estimation also affects to the angular descriptor.

Figures 4.14 and 4.15 respectively show examples of good and wrong matches in the supermarket. In the wrong matches figure 4.15, a woman wearing a brown coat with black trousers holding a brown bag is mismatched with a woman wearing similar clothes. In this case, the angular distribution of the color does not help too much. In the second match, the algorithm matches a man at the entrance wearing grey sweater with black coat with a man at the exit wearing a light brown shirt with dark brown coat. The reason of this match is not clear. Probably it is due to the similar appearance that the descriptors have after color normalisation.

### 4.5.3 Conclusions

In this section, we have introduced the 3D Bodyprint descriptor. This descriptor extends the Bodyprint descriptor presented in Section 4.4 by taking into account the angular information of the color features extracted from the raw feature descriptor. The 3D Bodyprints provide a good spatial representation of the person appearance by using the angular information of the color features.

The new descriptor has been evaluated the same scenarios as the Bodyprints. We have studied the influence of a characteristic hyper-parameter of this new descriptor, the $NBO$. In the school scenario, increasing the number of angular bins directly contribute in improving the re-identification probability. However, in the supermarket, which has uncontrolled conditions, increasing the number of angular bins does not help. The explanation of that behavior is directly related to the estimation of the trajectory of the person, which directly affects the descriptor.

Therefore, the optimal $NBO$ directly depends on the scenario. For those scenarios where the trajectories are kept constant, increasing $NBO$ may help, specially when the camera covers different angles of the person during their tracking. On other hand, in those challenging scenarios where the users often change their trajectories and the center of gravity cannot be accurately estimated, lower $NBO$ values would be recommended to avoid the shifting in the angular features.

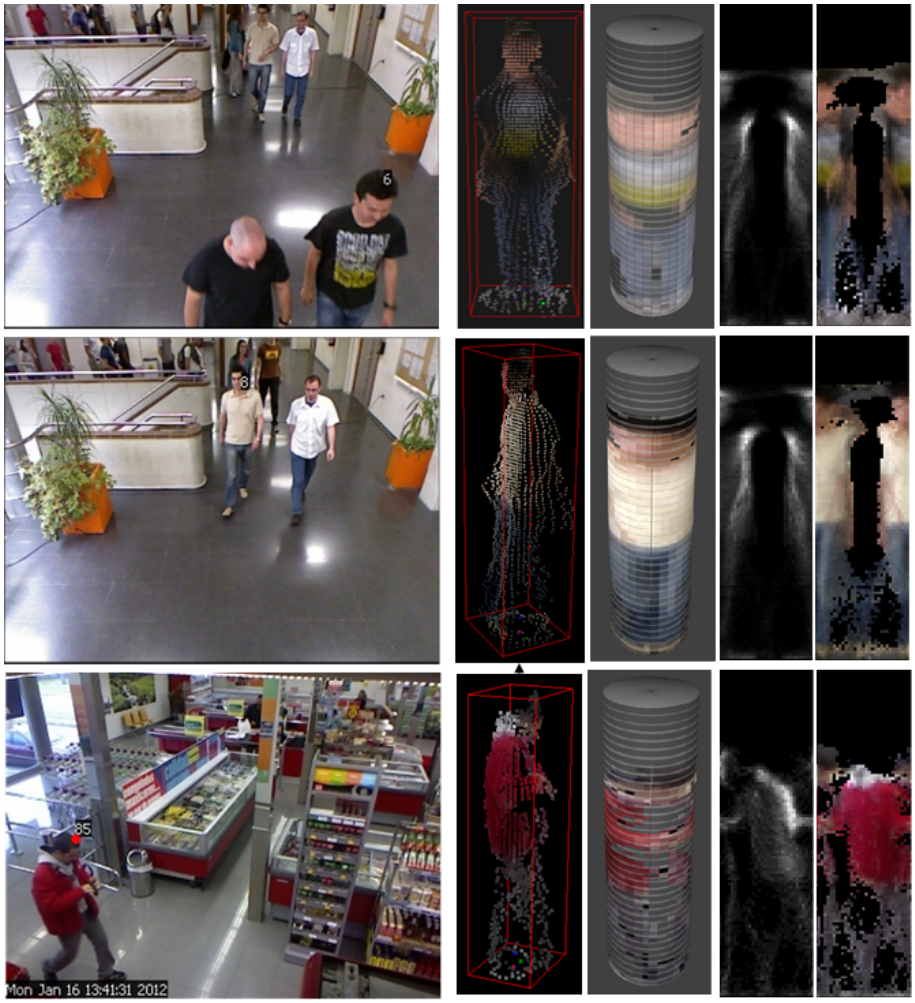



Figure 4.12: Examples of 3D Bodyprints for different people at different scenarios. The top and middle row show the school hall; the bottom row shows the supermarket. The first column shows the key image of the person during their tracking; the second column shows the raw features; the third column shows the 3D Bodyprint; the fourth and fifth columns respectively show the weights and colors of the 3D Bodyprint in an unrolled representation, where the vertical and horizontal axis represent the person height and angle.

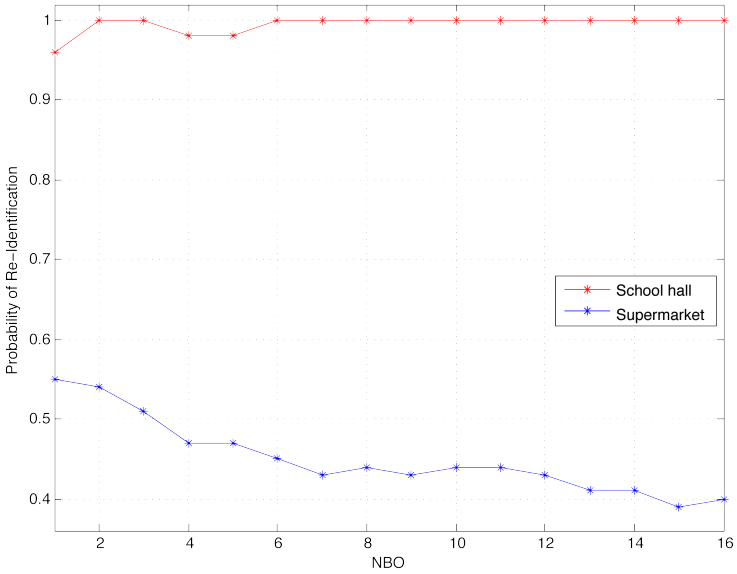



Figure 4.13: Probability of re-identification varying NBO.

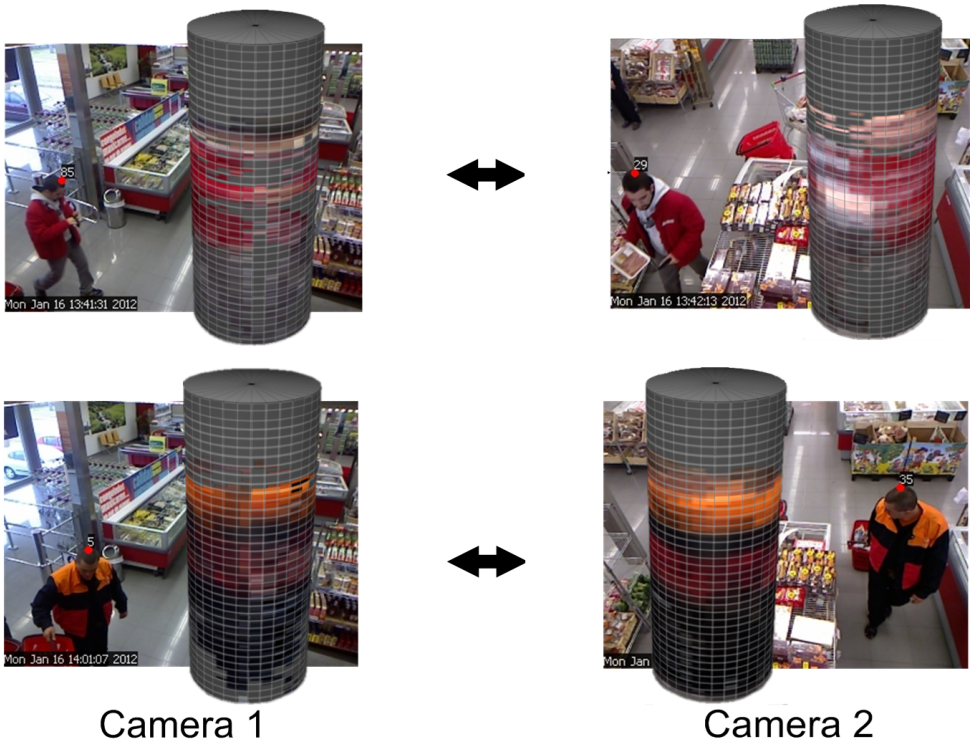



Figure 4.14: Examples of good matches in the supermarket scenario. On top of each key image, the 3D Bodyprint is depicted. The left column shows the key image at entrance; the right image shows the key image at the exit.


Figure 4.15: Examples of wrong matches in the supermarket scenario. On top of each key image, the 3D Bodyprint is depicted. The left column shows the key image at entrance; the right image shows the key image at the exit.

## 4.6 Bags of Appearances

The person descriptors presented so far (Bodyprints 4.4 and 3D Bodyprints 4.5) marginalise the time component of the raw features, which means that the resultant features represent the mean person appearance over time into one single vector. As we have demonstrated, both descriptors naturally remove small artifacts and outliers thanks to the temporal averaging, yielding good re-identification rates. However, they both basically fail in similar situations when the temporal signature is strongly affected by outliers, which drastically contribute in the time-averaged feature vector.

Outliers are appearance variations that can be introduced by different situations. For example, changes in the person pose because of the articulation of the human body. Also, small variations in height due to the gait. In re-identification systems, changes in illumination may strongly affect the person appearance during their tracking. Moreover, in this kind of difficult scenarios such as the supermarket one, where the people can behave naturally with no restrictions, the presence of undesired objects in the scene such as shopping cases or grabbed objects is a very common situation. In all these situations, calculating the average person appearance over time may generate strongly contaminated features because the color distribution of the pixels belonging to one bin is no longer monomodal.

In this section, a new descriptor that considers all the different person appearances over time is presented with the name of *Bag of Appearances - BoA*. A BoA is a container of color features that fully represents a person by collecting all their different appearances over time. Color features are 1D-feature vectors that are generated from each single frame of the raw feature vector by marginalysing in angle and radius. The aspect of each 1D-feature vector is the same as the Bodyprint, but in this case the feature vector represents the appearance at a particular time instant. Based on the results from the previous section, we decided not to use the angular information.

The key point in this representation relies in the fact that the person appearance is not averaged, so outliers are naturally removed in the matching process. Matching of bags is performed in a knn probabilistic framework by accumulating the probability of matching for all of the elements of each bag.

The following points summarise the points that will be covered in this section:

- Introduction of the BoA descriptor
- Description of the similarity metric using knn under a probabilistic framework
- Details of the algorithm training
- Algorithm assessment

### 4.6.1 Algorithm Description

#### 4.6.1.1 Bag of Appearances

The proposed technique, BoA, characterizes a person by collecting all the different person appearances during the tracking. This representation allows to model the person appearance from different points of view, poses and naturally deals with outliers. It is also indicated for those scenarios where the lightning conditions may vary during the person tracking.

Figure 4.16 illustrates this concept. The figure shows the matching of two different people at the supermarket scenario. The left part shows the person at the entrance. The right part of the figure shows the person at the exit. For each access point, the keyframe together with the Temporal Signature and the Bodyprint feature vector (see Chapter 4.4) are depicted. In the keyframe, the most representative person segmentation is shown during the person tracking to illustrate the different changes in point of view, pose and the outliers (Figure 4.17 provides finer detail of the different person appearances during the tracking at the entrance and exit point).

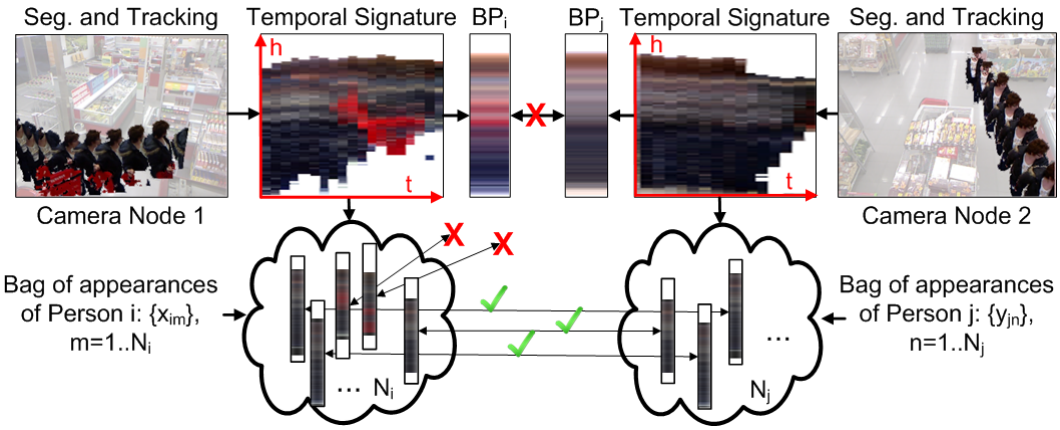



Figure 4.16: Example of the variability of the person appearance over time seen from the cameras at entrance and exit points in the supermarket

If we focus on the Bodyprint feature belonging to the person at the entrance point (Figure 4.16), we see that the mean colors of the regions belonging to the pelvis and legs are clearly affected by the influence of the shopping basket, which is taken from the pile and is connected to the body due to a bad segmentation. When both Bodyprint features are compared, the corrupted colors provoke a mismatch. However, if we focus on the Temporal Signature we can see that there are many uncorrupted frames in both sequences which could be used to produce a correct matching.

The concept of BoA is similar to the Bag of Features [Varma 2005] [Winn 2005] [Fei-Fei 2005]. However, fundamental differences arise due to the nature of the color features and the huge variability in the items representing each class. The

Figure 4.17: Detail of the person segmentation at each frame during the tracking. The top row shows the segmentation at the entrance point. It can be seen that the person changes her moving direction and therefore the appearance changes. Also, she takes the shopping basket from the pile and the segmentation algorithm takes the basket as a part of the person. The bottom row shows the segmentation at the exit point. In this case the person maintains the perspective towards the camera.

main difference is that we use separate bags for each person and do not perform global clustering. Although clustering of the appearances in each bag is a promising idea to reduce the computational cost of the algorithm, we decided not to use it in this work and to focus only on how multi-instance representation affects to the re-identification rate.

Since re-identification is performed across different cameras, it is very important to compensate for illumination variations. To do so, we follow the same approach as in Bodyprints. The brightness of the i-th person, $\mathcal{M}_i$, is obtained as the weighted mean of all the R, G, B pixel values of a person $\mathcal{P}_i$, and is subtracted from the color components of the appearance vectors.

#### 4.6.1.2 Probabilistic Framework for BoA matching

Suppose that we want to find the person $X_i$ from camera A among the set of $N_B$ persons from camera B denoted as: $\{Y_j\}_{j=1..N_B}$. If person $X_i$ is represented by his bag of appearances, $\{\mathbf{x}_{im}\}_{m=1..N_i}$, for each $\mathbf{x}_{im}$ we find its k-nearest neighbors among all the items from all the bags of camera B $\{\mathbf{y}_{jn}\}_{j=1..N_B,n=1..N_j}$. The comparison between appearance vectors has been carried out using the weighted correlation B.2, which allows to deal with missing data. The confidence provided by the k nearest neighbors of $\mathbf{x}_{im}$ is accumulated in a histogram $H_i$. Each bin of $H_i$ corresponds to a different person of camera B. This cumulative strategy makes the measure more robust than a multiplicative approach, where the lower values penalize the other. In order to accumulate the information about the confidence of each element, we convert the correlation measures into posterior probabilities $P_c$. The posterior probabilities $P_c$ are estimated from the correlation likelihoods $L_c$ using the Bayes rule 4.10.

As mentioned above, our weighted correlation, used to compare frame appear-

ances, relies only on the common visible parts of both appearance vectors to deal with missing data (oclusions). Unfortunately, this method discards the person's height information which is also very discriminant. In order to enhance the probability of matching, we obtain the difference of the average heights between pairs of persons' tracks $\Delta h$ .

This difference is also transformed into posterior probability $P_h$ that measures the probability that two tracks belong to the same person given $\Delta h$. Since both, appearance and height information, are converted into probabilities and they are independent, which is a reasonable assumption, we multiply their values to fuse their information.

The equation 4.10 shows the posterior probabilities for the correlation and $\Delta h$, together with the combination of both, $P_{match}$, which determines the probability of matching given two feature vectors belonging to different bags.

$$\begin{aligned} P_c &\propto Likelihood_c * Prior_c \\ P_h &\propto Likelihood_h * Prior_h \\ P_{match} &= P_c * P_h \end{aligned} \tag{4.10}$$

The overall process of matching can be seen in the self-explanatory pseudocode at Algorithm 1.

### 4.6.2 Evaluation

In this section we analyse the algorithm performance by calculating the probability of re-identification for the two datasets. To that end, we identify two stages: training and testing.

In the training stage, the Likelihood functions that model the correlation and height difference need to be calculated. Also, the hyperparameters of the knn matching strategy need to be set up. The supermarket training dataset has been used in these experiments.

In the testing stage, we use the parameters found during the training to test the system performance on the school hall and supermarket dataset. The re-identification probabilities have been calculated.

#### 4.6.2.1 Training

One key aspect in the matching of BoA features is that the confidence scores obtained from feature comparison are transformed into posterior probabilities. To do that, we model the intra-class and inter-class distributions. These Conditional Probability Density Functions - PDF - are the Likelihood functions of the classes. Figure 4.18 shows the PDF for correlation and $\Delta h$ variables.

The top image from Figure 4.18 shows the Likelihood distribution $L_c$ of the distance between $\mathbf{x}_{im}$ and $\mathbf{y}_{jn}$ for the match and mismatch cases. The bottom image from the same figure shows the Likelihood distribution $L_h$ of the height

**Definition**
$X$ := People in node A;
$Y$ := People in node B;
$h_X$ := Heights of people in node A;
$h_Y$ := Heights of people in node B;
$C_i$ := Number different people in the datasets;
**Dataset normalisation**
Remove brightness variations of each person independently in nodes A and B;
**foreach** *person* $X_i$ *in Node A* **do**
  Define $H_i$ as the class histogram that represents the probability of matching person $X_i$ with dataset $Y$;
  **foreach** *appearance* $\mathbf{x}_{im}$ **do**
    % Find $k$ nearest neighb. of $\mathbf{x}_{im}$ in cam. $B$;
    **foreach** $\mathbf{y}_{jn} \in \{knn\_of(\mathbf{x}_{im})\}$ **do**
      % Convert distance to probability
      $P_a \leftarrow Dist(\mathbf{x}_{im}, \mathbf{y}_{jn})$
      % Calculate height difference between persons $X_i$ and $Y_j$;
      $\Delta h_{ij} = abs(h_{X_i} - h_{Y_j})$;
      % Transform $\Delta h_{ij}$ to probability;
      $P_h \leftarrow \Delta h_{ij}$;
      % Add observation to probability histogram;
      $H_i(j) = H_i(j) + P_a.P_h$;
    **end**
  **end**
  % Find person match;
  $match = argmax(H_i)$;
**end**

**Algorithm 1:** People matching using bag of appearances

difference $\Delta h$ for the match and mismatch cases. For simplicity, Likelihoods are modeled as Normal distributions in all cases.

First, we focus on the extraction of the Gaussian parameters that model the correlation distribution of the same-class matches and different-class matches. We model the *match* distribution by the Normal $N(0.63, 0.19)$, and the *mismatch* distribution by $N(0.13, 0.32)$, obtaining a class separability of $S = 1.6481$, where $S$ represents the Fisher Linear Discriminant [Fisher 1936].

Second, we focus on the extraction of the Gaussian distribution parameters for the case of difference of heights $\Delta h$, where we model the *match* and *mismatch* distributions by $N(0.60, 1.89)$ and $N(0.55, 27.71)$ respectively.

Third, we seek the optimal number of neighbours, $k_{opt}$, for the knn algorithm by using the training dataset applying cross-validation to avoid over-fitting. The dataset has been divided into 7 different subsets, where 6 have been used for training

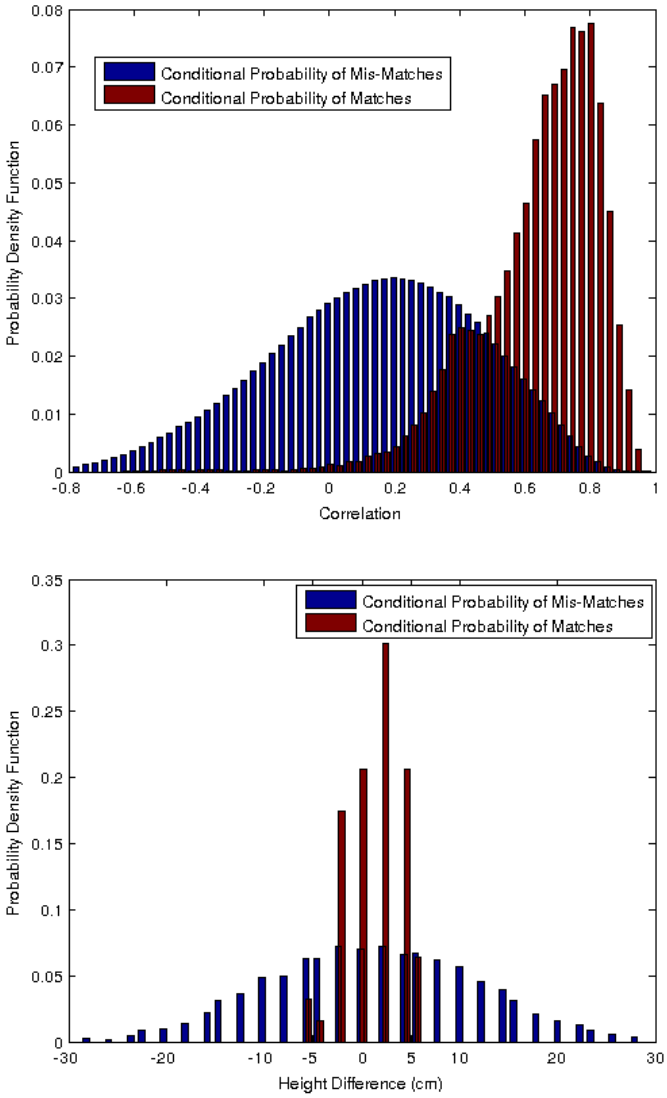



Figure 4.18: Likelihood distributions. The top image shows the Likelihood distribution of the appearance correlation values for the match and mismatch cases. The bottom image shows the Likelihood distribution for the height differences for the match and mismatch cases.

and 1 for testing. The test subset is subsequently swapped with any of the other used from training. The process is repeated 7 times and the mean probability of re-identification is calculated. The maximum probability of re-identification $P_R$ varying the number of neighbors from 1 to 10 is $P_R = 43.5$, found for $k_{opt} = 2$. Table 4.6.2.1 shows the variation of $P_R$ varying the number of neighbors from 1 to 10.

Table 4.4: Analysis of the optimal number of neighbours.

| k | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $P_R$ | 40.3 | 43.5 | 41.9 | 41.9 | 38.7 | 38.8 | 37.1 | 37.1 | 37.1 | 35.5 |

#### 4.6.2.2 Testing

After tunning the parameters of the knn algorithm and modeling the correlation and heights into a probabilistic framework, we evaluate the *BoA* matching approach using the school and supermarket testing datasets.

In the school hall, we get $P_R = 100\%$. Note that this database is more simple than the supermarket one. People walk naturally, they do not grab objects, and are always moving in the same direction.

In the supermarket, we obtain a probability of re-identification of 58.11%, which clearly outperforms the re-identification rates obtained using Bodyprints for the same dataset, which was of 53.05%. Figure 4.19 shows some challenging people that motivate the use of the BoA strategy since their temporal appearance presents significant variance. Note that for all these cases, the mean colors retrieved by Bodyprints would not be fully representative of each person. For the man with label 23 pushing a trolley, there is a period of time where his hand is segmented correctly and is connected to the body, but there is also a period where the hand does not appear. The women with labels 50, 43 and 28 are taking a shopping basket from the pile. The inaccurate segmentation yields a red color during several frames of the temporal signature. On other hand, both the women with labels 28 and 3 hold a white paper at the end of the signature. Finally, the man with label 14 shows a problem of illumination variability and missing data together with the contribution of the shopping basket.

The reason of the improvement achieved by BoA relies on the independent treatment of the temporal information of the appearance. Figure 4.20 shows the accumulated probability of matching for two different people in the supermarket. The sparsity of the histogram of person with ID 28 shows that several people presented similar appearance to the person of interest during small instants, mainly at the central and final frames of the track. However, the contribution of the first frames, where the appearance is not corrupted yet, are essential for providing the correct match of this person.

The histogram of person 43 shows an easier example of matching, although most of the first frames are badly classified.

On other hand, the Figure 4.21 shows an example where the matching is correct but the confidence is too low. Tracks with ID 20, 23 and 24 are really close to the correct one (ID 14). Figure 4.22 shows the key frame and the Temporal Signature of the most probable people. There are several reasons that explain the high number of candidates. First, the query image presents too many missing pixels in the lower body which causes that the matching is only performed based on the upper


Figure 4.19: Examples of challenging tracks.

part. Moreover, the upper part is black and few distinctive and also presents quite unstable illumination, which causes that this person has different appearances, all of them with significant number of frames. Therefore, although the real match (see Figure 4.22 top row) presents a very long temporal signature with many different appearances, most of them do not correlate with those of the query. Other people like those with ID 20, 23 and 24 do also have similar appearances. For example, we see that the track 24 correlates with the first frames of the query image. The reason is that the predominant color in these frames is the blue. In the other target images, the dominant color in the upper body is the black or dark blue.

### 4.6.3 Conclusions

In this section we have introduced the concept of *BoA*. This new descriptor method follows a Bag-of-words fashion in the sense that a bag describing the person appearance is formed from all the different person appearances extracted from the tracking. Each item in the bag represents a different person appearance generated from the color information at a single frame. The matching of bags is performed using knn strategy under a probabilistic framework.

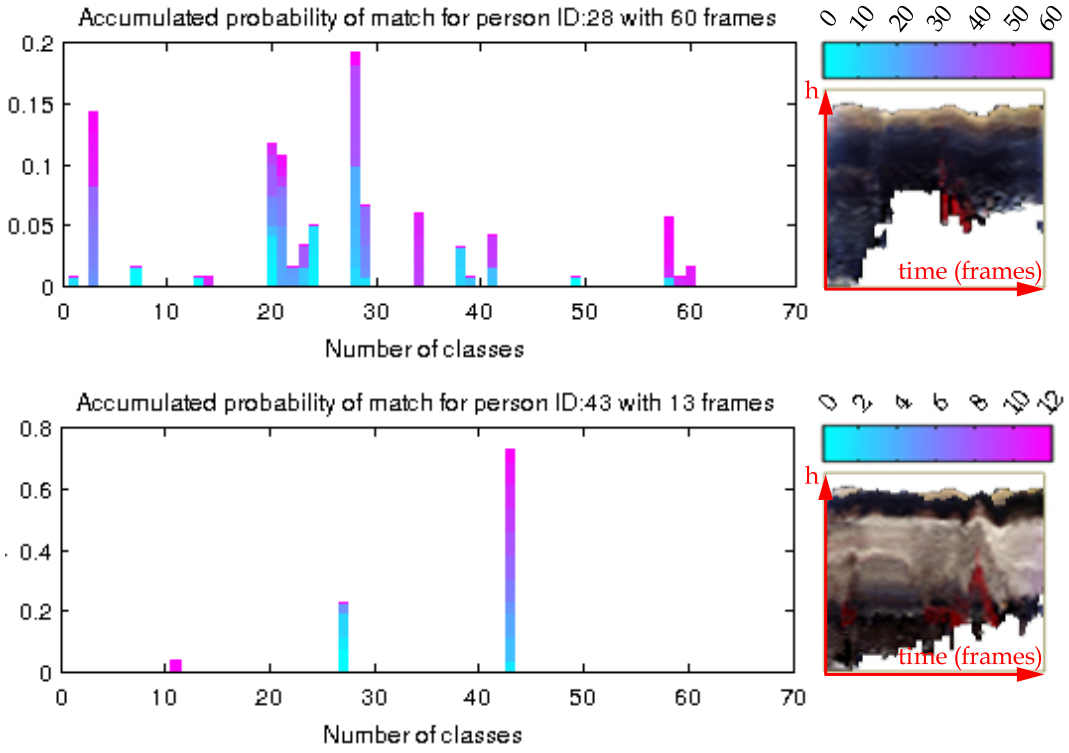



Figure 4.20: Accumulated probability-of-matching of a person at the entrance against 73 different people at exit. Two different people are depicted on top and bottom rows, with IDs 28 and 43 respectively. Left column represents the histogram of the matching probabilities using knn with 2 neighbors. The right column shows the Temporal Signature of the person at the entrance point. On top of the temporal signature image, there is a color bar that associates each frame with a different color for better interpretation in the histogram image.

The experiments with BoA demonstrate that this new method naturally copes with outliers. Artifacts like trolleys or grabbed objects do not contribute negatively in the matching process. Generally speaking, this method is indicated for situations where high number of outliers may appear.

Another interesting property of BoA is that this method may naturally capture the 3D person appearance when the person is seen from different perspectives during the tracking. In these situations, the method may gather a complete representation of the person appearance.

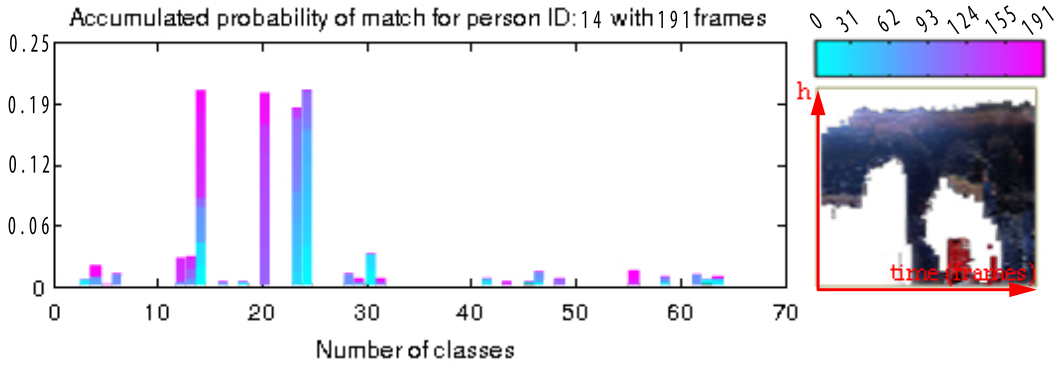



Figure 4.21: Accumulated probability-of-matching of a person at the entrance with ID 14 against 73 different people gathered at exit. The correct match is the bin 14. However, there are other bins with similar scores. To understand the reason for such similarities, the Temporal Signatures of the most probable tracks (14, 20, 23 and 24) are given in Figure 4.22.


Figure 4.22: This figure shows the most probable matches at the exit point for the query person with ID 14, as shown in Figure 4.21. The left column shows the key image. The right column shows the Temporal Signature for each person. The top row (id 14) shows the correct match.

## 4.7 Latent Features

The already introduced description methods present different approaches to deal with the high-dimensionality of the raw-features by means of data simplification. Bodyprints 4.4 remove the temporal and angular dependence. 3D Bodyprints 4.5 remove the temporal dependence but preserve the angular information of the color features. BoA 4.6 exploits the temporal information and removes the angular dependence. All these approaches have demonstrated to yield good performance in the school and supermarket scenarios. However, there are yet some points that can be further exploited to increase their performance. Let's analyse the methods in order to identify those points.

The goodness of Bodyprints relies in the fact that the information is averaged over time and angle, which easily captures the main appearance of the person into a compact representation. However, its quality may be degraded in complex scenarios by some environmental variables, such as non-uniformly distributed changes in illumination, the presence of carried objects, occlusions or high feature correlation due to flat-color clothes.

Figure 4.23 shows some challenging examples. A representative frame of the tracked person is shown next to the temporal signature and its corresponding Bodyprint in each example.

The first column of the Bodyprint represents the temporal mean color at each height and the second column represents the variance of the three color channels. In Figure 4.23 (a), a person enters the store and gets a red shopping basket from a pile. In the temporal signature, the shopping basket appears as an outlier that increases the red color variance specially at the lower part, as shown in Figure 4.24. Other carried objects, as the child shown in Figure 4.23 (b) or shop items as in Figure 4.23 (c) and Figure 4.23 (d), also produce outliers in the Bodyprint and increase the color variance at the corresponding heights. Note that they should be treated as outliers because the person appearance is different from entrance to exit. Finally, in Figure 4.23 (c) the feet of the person are never visible, so that the corresponding parts of the Bodyprint are missing.

3D Bodyprints inherit the properties of Bodyprints, but take into consideration the angular dimension in the feature vector. This descriptor is specially indicated for those scenarios where the people can change their trajectory. Apart from the advantages and disadvantages inherited from Bodyprints, the 3D Bodyprints may have problems with the estimation of the center of mass and the trajectory vector, which cause a phase shifting in the color information and affects the matching.

BoA approach copes very well with the presence of outlying elements in the scene since the person appearance is taken and matched independently at each frame. However, the appearance at each frame may be noisy due to the gait movement, illumination, occlusion or sensor accuracy. Figure 4.23 (c) shows the

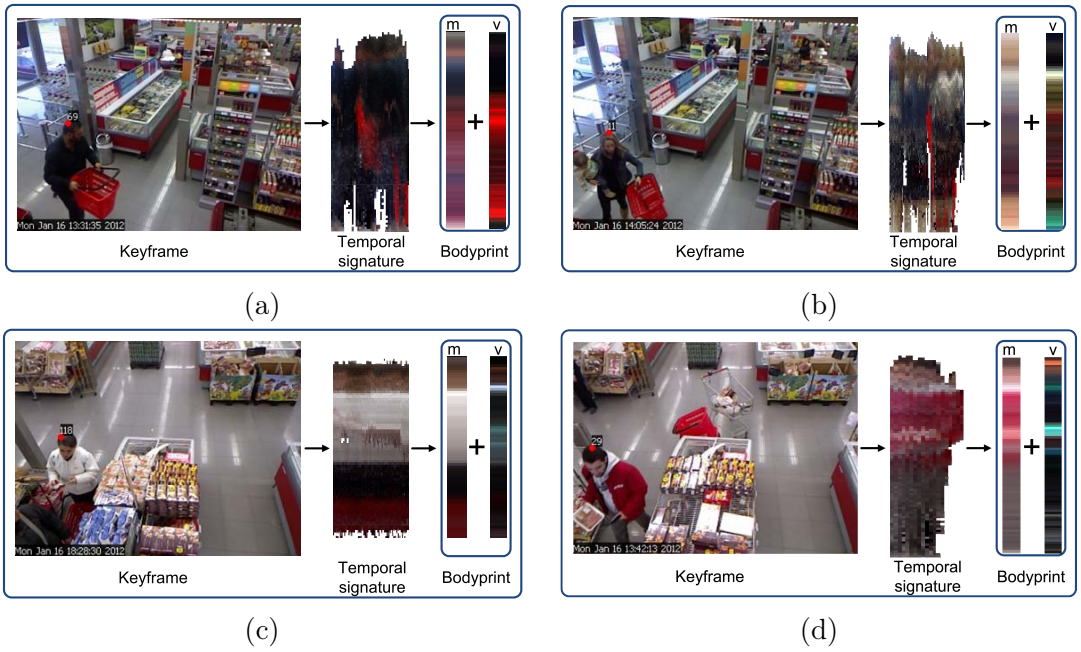



Figure 4.23: Some examples of Bodyprints and BoA descriptors with outliers generated by carried objects. In addition, figure a and c present missing data; figure a and b are strongly affected by the occlusions, where the resultant Bodyprints do not faithfully describe the appearance; figure d shows a small outlier at the person hand due to a carried object.

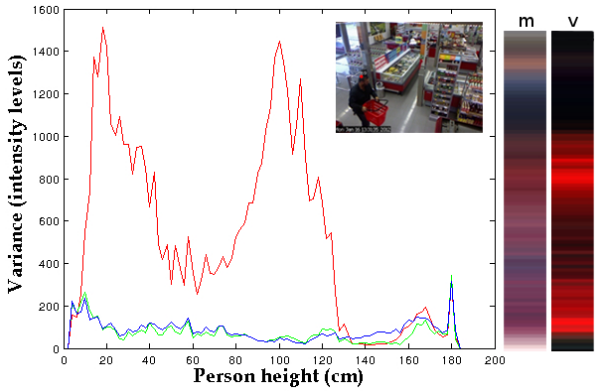



Figure 4.24: Detail of the color variance of a contaminated Bodyprint (Figure 4.23 (a)). The appearance of the red shopping basket produces high variance in the red channel.

effect of gait movement, which produces small variations in the color information in the vertical axis. Figure 4.23 (b) shows the Temporal Signature, which has been affected by occlusions. When comparing each frame independently, there are a high number of frames that are incomplete and would not provide high matching scores.

In addition to all the above mentioned challenges, other intrinsic parameters such as the height of each individual stripe, which generates high dimensional and correlated feature vectors, may also affect the quality of the descriptor vectors.

Table 4.5 summarasies several of these factors and how they affect the descriptors.

Table 4.5: Examples of ambiental factors and how they affect the Bodyprints. For clarity, the acronym *n.u.* refers to the words *non-uniform*.

| Factors | Bodyprints | 3D Bodyprints | BoA |
|---|---|---|---|
| N.u. light changes | light<br>*n.u. color var* | light<br>*n.u. color var* | light<br>*n.u. color var* |
| Pose variations | severe<br>*outliers* | severe<br>*outliers* | light |
| Gait | light | light | severe<br>*noisy observation* |
| Trajectory estimation | NA | severe<br>*color shifting* | NA |
| Occlusions | light | light | severe<br>*missing data* |
| Carried objects | severe<br>*outliers* | severe<br>*outliers* | light |
| Flat-colored clothes | severe<br>*feature correl* | severe<br>*feature correl* | severe<br>*feature correl* |

All in all, the proposed methods present common challenges, mostly related to the strong feature correlation, the presence of outliers and missing data. In this section we propose a set of probabilistic-based dimensionality reduction techniques to handle these problems. In order to evaluate these techniques, the Bodyprint descriptor will be used in the supermarket database. Bodyprints are selected because they are a good representation of all the proposed methods. The supermarket is chosen because it is a challenging database that gathers all the problems listed above. Nevertheless, the approach discussed in this section can be applied to all of the methods and scenarios presented in this Thesis.

### 4.7.1 Algorithm Description

In this section, we present a set of statistical models to address the problem of dimensionality reduction under a probabilistic framework. These models transform the Bodyprint features into a new set of features, named *Latent Features*, $\mathcal{L}$, which are represented in a new coordinate system with lower dimensionality. In this new space, the feature correlation is removed and the noise that does not lie in the reduced feature space is also removed, therefore alleviating the problems caused by outliers.

The process of dimensionality reduction basically consists in finding the projection matrix, $\mathbf{W}$, which transforms a Bodyprint vector into the *latent features*. The axis of this new coordinate system are set to capture the maximum possible variance of the original observations. These axis, which are orthogonal, are called the *Principal Components*. Therefore, the *latent features* are defined as follows:

$$\mathcal{L} = \overline{RGB}' \times \mathbf{W} \tag{4.11}$$

where $\mathcal{L}$ represents the *latent features* and $\overline{RGB}'$ represents the Bodyprint color features. Each column of $\mathbf{W}$ represents a principal component, the components are ordered so that the first principal component gathers the maximum data variance, and the maximum number of components is equal to the number of observations, which is the total number of people in the training set.

The projection matrix is calculated in a training stage using the training dataset under cross-validation. In classical problems, this matrix can be calculated using PCA or LDA techniques. However, they can not cope with missing data and do not consider additional information such as the variance of the observed features, which provides useful information about the confidence of the observed values. For this reason we propose to apply probabilistic latent variable models, which are equivalent to classical techniques but cope with missing data and also allow to incorporate information about the variance of the observed data.

Figure 4.25 shows the complete processing pipeline, where the *latent features* are extracted from the Bodyprints. In the sequel, matrices are represented as bold upper case, column vectors as bold lower case and real scalars as italic lower case.

#### 4.7.1.1 Latent variable models

Probabilistic latent variable models are usually represented by graphs [Bishop 2006], as shown in Figure 4.26. In the graph, model variables are placed in circles. The difference between white and blue circles is that in the latter case the variables are observed or measured, where the other variables are not observed and can be inferred from observed features and for this reason are called *latent features*. The rectangular box surrounding the circular variables in the graph represents a plate.

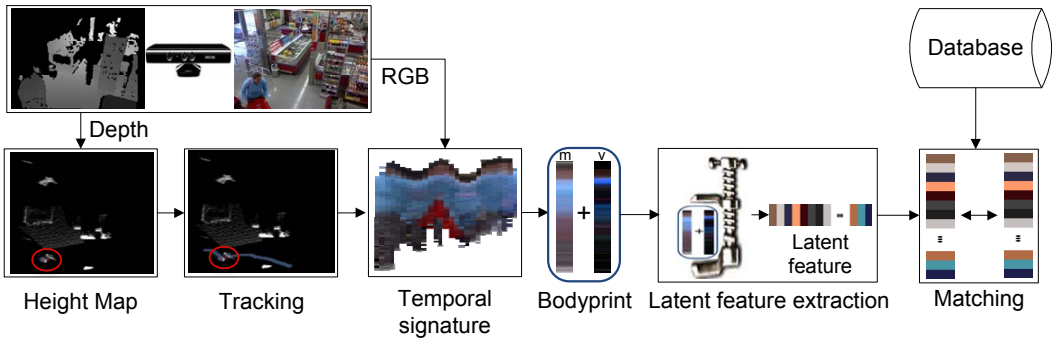



Figure 4.25: Latent feature extraction for a single node in the camera network. The detected and tracked person is highlighted using a red circle in the Heigh Map and Tracking images. The Bodyprint features are used as input features. The resultant features, named *latent features*, are directly used in the matching process.

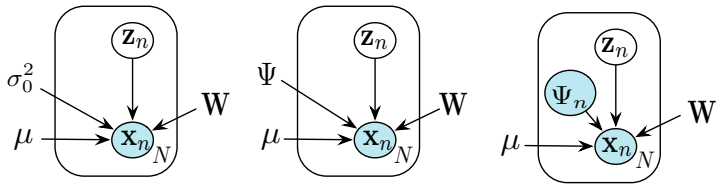



Figure 4.26: Different latent-feature extraction models. In a) the noise is considered isotropic and constant for all samples. In b) the noise is modeled by a diagonal matrix and is constant for all samples. In c) there is a different diagonal matrix representing the noise for each sample.

Each plate embeds $N$ training samples of which only a particular $\{\mathbf{x_n}, \mathbf{z_n}\}$ pair is depicted. Both $\mathbf{x_n}$ and $\mathbf{z_n}$ are vectors of length $l_x$ and $l_z$ respectively. While $\mathbf{x_n}$ represents the observed color features of a person (Bodyprints), $\mathbf{z_n}$ denotes its corresponding latent variables ($l_z < l_x$) that we want to extract by inference. Terms outside the plate indicate the model parameters: mean $\boldsymbol{\mu}$, projection matrix $\mathbf{W}$ and noise $\sigma$ that are shared among all training and test samples. These parameters are obtained during the training stage using the EM algorithm [Roweis 1998] because for some of the proposed models do not exist a maximum likelihood closed solution.

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

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

$$\mathbf{x_n} = \mathbf{W}\mathbf{z_n} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{4.12}$$

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

It can be demonstrated [Bishop 2006] that in PPCA, the posterior distribution of the *latent features* can be expressed as:

$$p(\mathbf{z_n}|\mathbf{x_n}) = N(\mathbf{z_n}|\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x_n} - \boldsymbol{\mu}), {\sigma_0}^2\mathbf{M}^{-1}) \tag{4.13}$$

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

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

$$\begin{aligned} \mathbb{E}[\mathbf{z_n}] &= \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x_n} - \boldsymbol{\mu}) \\ \mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T] &= \sigma_0^2\mathbf{M}^{-1} + \mathbb{E}[\mathbf{z_n}]\mathbb{E}[\mathbf{z_n}]^T \end{aligned} \tag{4.14}$$

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

$$\begin{aligned} \mathbf{W} &= \left[\sum_{n=1}^{N}(\mathbf{x_n} - \boldsymbol{\mu})\mathbb{E}[\mathbf{z_n}]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T]\right]^{-1} \\ \sigma_0^2 &= \frac{1}{ND}\sum_{n=1}^{N}\{\|\mathbf{x_n} - \boldsymbol{\mu}\|^2 - 2\mathbb{E}[\mathbf{z_n}]^T\mathbf{W}^T(\mathbf{x_n} - \boldsymbol{\mu}) \\ &\qquad + Tr(\mathbb{E}[\mathbf{z_n}{\mathbf{z_n}}^T]\mathbf{W}^T\mathbf{W})\} \end{aligned} \tag{4.15}$$

This process is repeated until convergence of the parameters. To evaluate the convergence, the EM algorithm uses the log-likelihood of the observed data as the objective function (see [Verbeek 2009] for details).

**PPCA with missing data.** The problem of the EM algorithm as formulated in Equations 4.14 and 4.15 is that it can not deal with missing values in the observed Bodyprints (caused by occlusions). Fortunately, the EM provides a natural way to handle missing values. The three main differences that are found when dealing with missing data are:

- The mean of the observed Bodyprints, $\boldsymbol{\mu}$, cannot be computed in a closed form and needs to be estimated in each iteration.
- The covariance matrix of the posterior distribution of the *latent features*, (Eq. 4.13), is different for each training sample as it depends on which variables of the Bodyprint are observed in each sample.

- The formulation to obtain the transformation matrix $\mathbf{W}$ is more complex because, in this case, each row of $\mathbf{W}$ needs to be calculated independently.

To indicate which Bodyprint features have been observed, we use the binary matrix $\mathbf{O}$, so that $\mathbf{O}(m,n) = 1$ if the $m$ feature of the training sample $n$ is observed. Similarly, $O_n$ is the set of indexes of the observed Bodyprint features for sample $n$ and $O_m$ is the set of samples for which feature $m$ has been observed. Using an element-wise formulation of Eqs 4.14 and 4.15, the EM steps can modified to deal with missing data using only the columns of $\mathbf{W}$ and rows of $\mathbf{x_n}$ that correspond to the observed values (see [Ilin 2010] for a detailed explanation). In the Expectation step the latent features for each sample are estimated using the observed data and the current estimates of the parameters:

$$\mathbb{E}[\mathbf{z_n}] = \mathbf{M_n^{-1}} \sum_{m \in O_n} (\mathbf{x_n}(m) - \boldsymbol{\mu}(m))\mathbf{w_m} \tag{4.16}$$

$$\mathbb{E}[\mathbf{z_n}\mathbf{z_n}^T] = \sum_{m \in O_n} (\sigma_0^2 \mathbf{M_n^{-1}} + \mathbb{E}[\mathbf{z_n}]\mathbb{E}[\mathbf{z_n}]^T) \tag{4.17}$$

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

$$\boldsymbol{\mu}(m) = \frac{1}{|O_m|} \sum_{n \in O_m} (\mathbf{x_n}(m) - \mathbf{w_m^T}\mathbb{E}[\mathbf{z_n}]) \tag{4.18}$$

$$\begin{aligned} \mathbf{w_m^T} \quad = \quad & \left[ \sum_{n \in O_m} (\mathbf{x_n}(m) - \boldsymbol{\mu}(m))\mathbb{E}[\mathbf{z_n}]^T \right] \cdot \\ & \left[ \sum_{n \in O_m} \mathbb{E}[\mathbf{z_n}\mathbf{z_n}^T] \right]^{-1} \end{aligned} \tag{4.19}$$

$$\begin{aligned} \sigma_0^2 \quad = \quad & \frac{1}{N} \sum_{m,n \in O} \{(\mathbf{x_n}(m) - \mathbf{w_m^T}\mathbb{E}[\mathbf{z_n}] - \boldsymbol{\mu}(m))^2 + \\ & + \mathbf{w_m^T}\mathbf{M_n}^{-1}\mathbf{w_m}\} \end{aligned} \tag{4.20}$$

**Factor Analysis (FA).** Factor Analysis is a latent variable model similar to PPCA. The main difference between them is that in the generative model of Eq. 4.12 the noise distribution in FA is:

$$p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon}|0, \boldsymbol{\Psi}) \tag{4.21}$$

where $\boldsymbol{\Psi}$ is a general diagonal $l_x \times l_x$ matrix. The advantage of FA compared to PPCA is that it is a more flexible model that can capture different noise levels

in the observed Bodyprint features. The example of Figure 4.24 shows that this situation can often appear in our re-identification context. Since $\mathbf{\Psi}$ is diagonal matrix, the noise is also considered independent for each Bodyprint feature in FA. This assumption is necessary to reduce the number of model parameters and avoid over-fitting.

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

$$\mathbb{E}[\mathbf{z_n}] = \mathbf{G_n} \sum_{m \in O_n} \mathbf{w_m} \mathbf{\Psi}_m^{-1} (\mathbf{x_n}(\mathbf{m}) - \boldsymbol{\mu}(m)) \tag{4.22}$$

$$\mathbb{E}[\mathbf{z_n} \mathbf{z_n^T}] = \sum_{n \in O_m} (\mathbf{G_n} + \mathbb{E}[\mathbf{z_n}] \mathbb{E}[\mathbf{z_n}]^T) \tag{4.23}$$

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

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

$$\begin{aligned} \mathbf{\Psi}_m = diag\{ & S_m - \\ & - \mathbf{w_m} \frac{1}{|O_m|} \sum_{n \in O_m} \mathbb{E}[\mathbf{z_n}] (\mathbf{x_n}(m) - \boldsymbol{\mu}(m))^T \} \end{aligned} \tag{4.24}$$

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

**Factor Analysis with known noise** . In Fig 4.26.b, it is shown the graph representation of FA. It can be noticed that $\mathbf{\Psi}$ remains constant for all samples and it is inferred from the training samples as the other model parameters. However, this can be a strong simplification because not all samples are equally corrupted by the noise. In this work, we propose a small modification to the FA model that is depicted in Figure 4.26.c. With this modification it is possible to introduce different noise distributions for each sample. The noise distribution for each sample is a zero-mean multivariate Gaussian with diagonal covariance matrix $\mathbf{\Psi}_n$. Notice that $\mathbf{\Psi}_n$ is placed in a blue circle what means that the noise distribution is measured and introduced into the model. The main advantage of this new approach is that introducing an estimate of $\mathbf{\Psi}_n$ we provide information about what features in each Bodyprint are more reliable and which color features are less important (higher color variance over time). Note that since $\mathbf{\Psi}_n$ is different for each Bodyprint, the information about the reliability of the features is different for each Bodyprint.

If the noise parameters are provided to the algorithm, the only model parameters that need to be estimated during the training stage are the model mean $\boldsymbol{\mu}$ and the projection matrix $\mathbf{W}$. The EM equations used to obtain the model parameters in

this case are the same as in the FA model except that $\mathbf{\Psi}_n$ no longer needs to be maximized (Eq. 4.24) and remains fixed (its value is provided as an input to the algorithm).

To obtain the noise distribution for each person, $\mathbf{\Psi}_n$, we use the corresponding temporal signatures (Figure 4.23). The idea is that for each height it is possible to extract the mean color over time together with the color variance since the color at each stripe may vary over time. Again, to reduce the number of parameters we assume that color features at different heights and color channels are independent ($\mathbf{\Psi}_n$ is diagonal).

#### 4.7.1.2 Model discussion

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

The experimentation in this section will discuss which of the different approaches is more convenient for the purpose of person re-identification in multi-shot scenarios, where people appearance over time can be described as a unique, stable color feature vector with known variance, such as the Bodyprint features. However, this approach can be extrapolated to be used with any of the other proposed methods, such as the 3D Bodyprints or BoA.

#### 4.7.1.3 Matching metric

In the Bodyprints, the height of each person is always quantized into 110 bins, where for each bin the average color information is stored. That produces that the feature vector automatically characterises the person height, provoking that the top stripes may be empty. In this section, the height of the people has been normalised, so all of them are equal. The idea behind of using a fixed number of color features in the Bodyprints is that we wanted to decouple appearance from height information. In preliminary experiments, we found that the height normalization of Bodyprints was very useful to reduce the dimensionality of the latent space because the variability of

the training samples is reduced with the normalization. However, height information is also very important for our re-identification objective and for this reason height information must be incorporated into the similarity metric.

Let $\mathbf{z}_i$ $\mathbf{z}_j$ be the *latent features* of person $i$ and $j$ respectively obtained using any of the methods proposed in Section 4.7.1.1. The similarity $\mathcal{S}(i,j)$ between these two people is defined as follows:

$$\mathcal{S}(i,j) = \mathcal{M}(\mathbf{z}_i, \mathbf{z}_j) e^{-\frac{\Delta H}{\nu}} \tag{4.25}$$

where the first term measures the appearance similarity using the *latent features* and the second term penalizes the difference in height, $\Delta H$, between the two people. The constant $\nu$ controls how the confidence decreases and its value was empirically determined ($\nu = 4$cm). In our experiments, several appearance measurements for $\mathcal{M}(\mathbf{z}_i, \mathbf{z}_j)$ are compared (see Section B).

### 4.7.2 Evaluation

In this section, we evaluate the proposed *latent feature* models in the context of people re-identification using Bodyprints. Information about the dataset on which experiments are carried is provided in A. To evaluate the performance of our system we use the probability of re-identification.

The proposed models (Figure 4.26) have been trained using the Bodyprints of the people at the entrance. Once the models are trained, the *latent features* of all Bodyprints are extracted. Finally, the matching is conducted using the metrics described in Section 4.7.1.3.

#### 4.7.2.1 Results

Similar to PCA, the columns of the projection matrix $W$ can be interpreted as the directions in the observed feature space that capture more variance of the training data. The maximum number of *latent features* that can be obtained is limited by the number of training samples in a dataset (in our case 63). Usually, latent features that capture more variance are considered more relevant since they contribute more to the reconstruction of the observed feature space. In most pattern recognition problems, the number of *latent features* to be considered is a trade off between capturing as much variance as possible and discard low variance features that capture the noise of the training set.

In this experimentation, we studied the influence of the number of *latent features* for all the models and metrics proposed in the section. As mentioned above the evaluation is carried out using the re-identification rates for each case. Figures 4.27, 4.28 and 4.29 show the re-identification rates for all the proposed cases. The first general conclusion is that a few *latent features* are enough to extract the relevant information in all the cases, what confirms our initial hypothesis about using dimensionality reduction techniques. Another result is that the Euclidean metric is the one that attains the best performance for the three latent models.

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

Figure 4.30 shows several correct matching examples using 20 PPCA *latent features*. The left and right columns show images obtained at the entrance and exit respectively. The persons that are matched in each case have been surrounded by a ellipse. Notice that in our dataset we do not impose any restriction on the person behaviour so in the second example the old woman is pushing a shopping trolley at the entrance but the cart does not appear at the exit. These examples show how our algorithm can effectively deal with big changes in pose and also with outliers produced by carried objects. For instance a shopping basket in the first example and a shop item in the third example. On the other hand, Figure 4.31 shows a few examples where re-identification failed. The examples clearly show the difficulty of our dataset where in many cases people wear similar clothes and re-identification is even difficult for the human eye.

In Table 4.6 we compare the re-identification rates achieved using the *latent features* with the Bodyprints 4.4. We see that the use of *latent features* have significantly raised the performance, which is mainly due to the fact that the *latent features* have naturally chosen the discriminative features that contain the relevant information.

Table 4.6: Summary of the best re-identification rates for the proposed methods

| Feature type | Re-identification Rates |
|---|---|
| Bodyprints 4.4 | 52.5% |
| PPCA | 62.5%@20 components |
| FA unknown variance | 61.5%@40 components |
| FA known variance | 62.5%@40 components |

### 4.7.3 Conclusions

In this section we introduced the concept of *latent features* for people re-identification. To generalise the problem, only the Bodyprint descriptors have been considered as input. Different probabilistic *latent feature* models have been proposed and compared in the analysis. It has been demonstrated that the use of these models minimizes some problems such as outliers, noise, missing data and correlation of

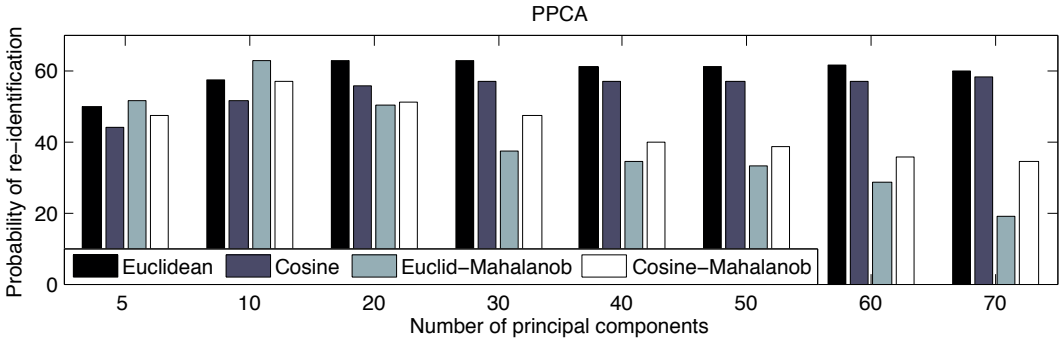



Figure 4.27: Probability of detection (rank $r = 1$) using PPCA method for different number of principal components.

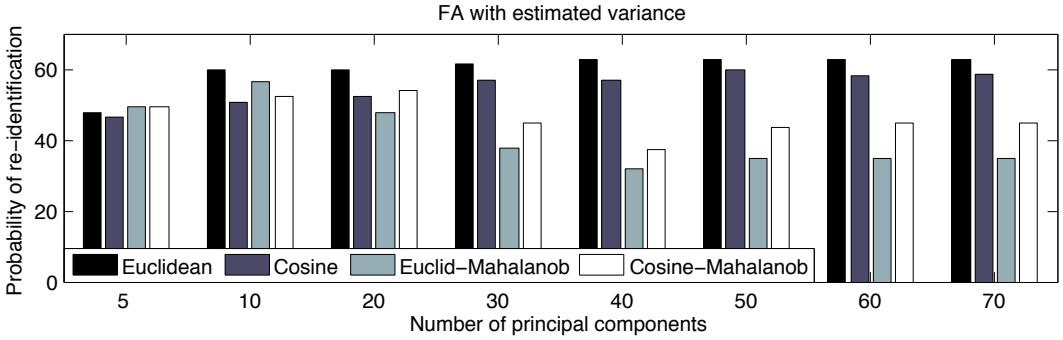



Figure 4.28: Probability of detection (rank $r = 1$) using FA with unknown variance for different number of principal components.

Bodyprint features.

The basic difference among the presented models is how noise is handled in each case. The results show that the best re-identification rates are obtained using PPCA, although FA with unknown variance provides more stable results since the re-identification rate does not decrease when more *latent features* are used. Compared to Bodyprints, the use of *latent features* significantly increases the global performance.

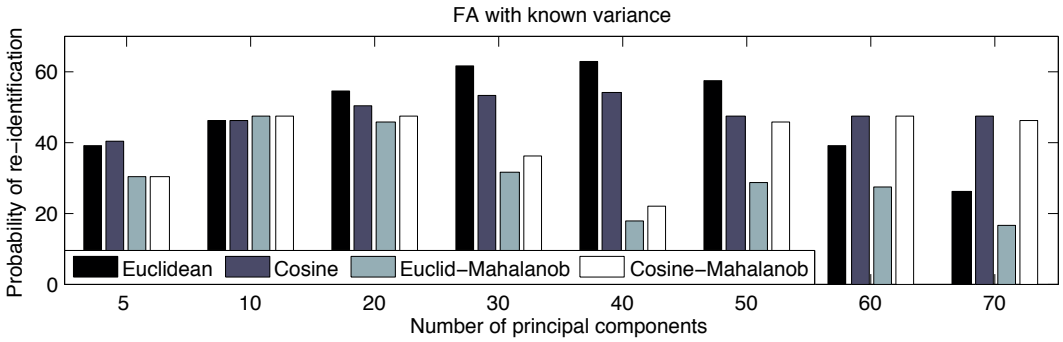



Figure 4.29: Probability of detection (rank $r = 1$) using FA with known variance for different number of principal components.


Figure 4.30: Examples of correct matches using PPCA method with 20 components (rank $r = 1$). Left and right columns show people at the entrance and exit of the shop respectively.


Figure 4.31: Examples of incorrect matches using PPCA method with 20 components (rank $r = 1$).

## 4.8 Comparison with the State of the Art

This section introduces the assessment of all the proposed person description methods. During the experimentation phase of this Thesis, no more works that used depth data for re-identification in similar conditions to those presented in our scenarios have been found. For example, the work described in [Barbosa 2012] uses depth information got from a depth sensor to perform the re-identification task. Features extracted with this method are generated using biometrical information, such as bones and joints, got from the person skeleton, which is extracted from the depth maps analysis. Given the nature of the descriptor, the authors claim that the users can change their clothes during the matching period. To assess their work, Barbosa et al have generated a public database containing video sequences of rgb and depth information got from a fixed camera with different recordings, where people change their clothes. To our knowledge, this database is the first one that includes rgb and depth information in a re-identification process.

Although this database is publicly available on the web of the authors, it could not be used in this work for comparison because the conditions of their database did not match our base hypothesis: people cannot change their clothes during the re-identification period. Moreover, the database from Barbosa et al does not change the camera view and require good image resolution in order to be able to retrieve the skeleton information. These restrictions are not taken in any the proposed descriptors in this Thesis, where the camera point of view can present a more vertical inclination, high resolution is not necessary and huge areas with different point of views are allowed. Therefore, the assessment of our method using Barbosa's database is not possible.

In order to carry out a fair comparison, given the fact that in the moment of this research none other compatible work was found that used rgb and depth information in the re-identification problem, the proposed person description methods are compared to some state-of-the-art methods that merely rely on color and texture information. A good representation of the state-of-the-art are the works presented by Farenzena et al [Farenzena 2010] and Zheng et al [Zheng 2011], described in Chapter 2.

Although these methods do not use depth information and therefore they are in disadvantage in the comparison, we just want to compare with them to show the real contribution of the Bodyprints in re-identification. In this section, a general comparison among the methods introduced in this Thesis (Bodyprints, 3D Bodyprints, BoA, Latent Features) and the state of the art methods (PRDC and SDALF) is conducted.

### 4.8.1 Experimentation Details

The following lines describe the most relevant parameter configuration for each method in the comparison. The best configuration of the methods introduced

in this Thesis has been previously discussed in this chapter. The parameters for PRDC and SDALF methods have been set according to the values that the authors recommend in their articles.

In all the methods, the training set has been used for parameter configuration. The test set has been used to evaluate the algorithm performance.

- Bodyprints: These features do not require complex parameter setting. The method naturally packs the Temporal Signature into a single vector. The Weighted Euclidean distance has been chosen for matching, where the features are based on color and the weights are represented by the number of pixels that contribute in calculating the mean color of a particular stripe. The exponential factor that considers the difference in the person heights has also been taken into account in the comparison formula.

- 3D Bodyprints: According to the results obtained in the Bodyprint evaluation section, the Number of Angular Bins $NBO$ is set to 2 . We use the Weighted Euclidean distance with the height factor as the similarity measure.

- BoA: The matching of the bags is carried out with trees with 2 neighbors. The weighted correlation is used as the similarity measure together with the height correlation.

- Latent Features: According to the previous section, the best *latent feature* model is the PPCA with 20 components. The Euclidean distance together with the person height difference are used in the matching function.

- PRDC: Using the training data, the distance function, which is the basis of this method, is learned based on the relative distance comparison [Zheng 2011]. This function is a quadratic Mahalanobis distance function.

- SDALF: According to their article [Farenzena 2010], the matching metric is made of a concatenation of three different measures: weighted Bhattacharyya distance for comparing the HSV histograms; the Euclidean distance for comparing the MSCR (Maximum Stable Color Regions) features; the Bhattacharyya distance for comparing RHSP (Recurrent High-Structured Patches). In order to find the matching function, we use our training set and use the following parameters -$\beta_{WH} = 0.4$, $\beta_{MSCR} = 0.4$, $\beta_{RHSP} = 0.2$ - that refer to the mentioned three different measures respectively.

For simplicity, the experimentation has been carried out using the supermarket database described in Appendix A. Note that this is the most challenging scenario,

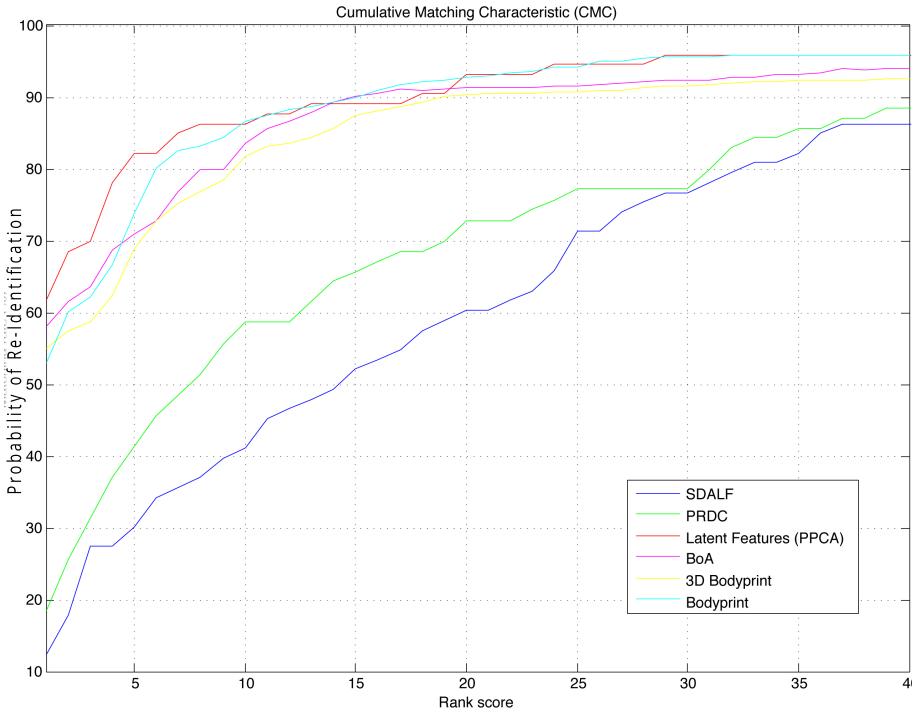



Figure 4.32: Methods performance comparison using CMC curves

which gathers real conditions with uncontrolled people behavior. CMC over 40 trials have been used to show the ranked matching rates. Note that none of the works used in the comparison use any other contextual cues such as temporal causality (a person at the exit must have entered first). Although this type of contextual information is very valuable because it reduces the effective search population in a real scenario, we preferred to ignore it so the results reflect only the re-identification abilities of the algorithms.

### 4.8.2 Evaluation

Figure 4.32 shows the performance comparison of the Bodyprints, 3D Bodyprints, BoA and *Latent Features* with the PRDC and SDALF reference methods. The *Latent Feature* approach overtakes all the other methods, even the reference methods, specially in the lower ranks in CMC curve. The BoA also provides good results. The Bodyprint descriptors (2D and 3D) present slightly worse performance than the other methods of their family, but yet overtake the reference methods, which show modest results.

In general, this remarkable difference in the results regarding the proposed methods in comparison to the state of the art is mainly due to the use of depth information that permits to describe people appearance considering their real height under a frontal perspective obtained by applying the inverse camera projection. Thanks

to the depth information, it is possible to accurately work with the spatial data to create more robust and reliable descriptors.

However, there is still a difference regarding the *Latent Features* combined with Bodyprints and the other methods of its family. This difference is so because this descriptor permits to handle missing data, removes noise, finds the discriminant features and provides a useful tool to work with the temporal signature of a person over time by compressing all the different appearances into a stable feature vector plus a variance that is absorbed by the latent model during training. Applying the PPCA to the other methods would also boost their performance drastically. Note that when applying this technique to the Bodyprint descriptor it has turned the most modest descriptor into the most reliable one.

Chapter 5

# Conclusions and Future Work

Contents



The problem of person re-identification using intelligent vision-based systems is an interesting topic that has become essential in domains such as surveillance, retail, marketing or sports. There is extensive work in the literature that address this problem in many different ways, most of them based on the color analysis from RGB images. However, problems such as variations in the person pose, people occlusion, the presence of carried objects, the use of similar clothes, changes in illumination or different camera geometries in the camera network, make this topic one of the most challenging in this field.

The recent introduction of depth sensors open a new paradigm in the image processing field. These sensors bring new possibilities by providing accurate depth images together with conventional RGB images at 30 fps. In this Thesis, we have used this novel technology to address the whole process of people re-identification from a new perspective. The key idea underlying this work is the use of depth information to accurately locate the color features in the 3D space. Based on this spatial representation of the color information, it is possible to develop new robust techniques that take into account the 3D shape and appearance of the people.

## 5.1 Conclusions

In this Thesis, we have defined a new network topology for person re-identification that uses rgb-depth cameras. Under this novel framework, we have elaborated a public rgb-depth dataset for person re-identification and we have proposed a set of family methods for feature extraction using depth information.

The public dataset contains two scenarios with different complexities and gathers 179 different people in total. This dataset was necessary in order to make the assessment of the proposed methods, since to our knowledge there was not any other public database for re-identification using rgb-depth data. This is a major contribution to society because researchers can use the dataset for running their tests and fairly comparing the methods performance.

In order to describe the people appearance, we first needed to define a strategy for person segmentation and tracking. Since depth information provides accurate estimation of the 3D position of the color pixels, we based the segmentation on the depth processing. Depth information at each frame was transformed into a Heightmap, which is a zenital representation of the scene. Using that representation of data, we easily segmented people by finding maximum points and then selecting all the connected pixels below that maxima. We demonstrated how this segmentation module naturally copes with occlusions and shows excellent precision and recall rates.

Once the accurate people segmentation was carried out, we introduced the description module. To ease the representation and the management of the features, we calibrated the scene to have the origin of the 3D coordinates onto the floor. In this 3D space, the data was represented by spatially distributing rgb values in a continuous domain, named pointclouds. In order to simplify the data, we set independent center of coordinates at the center of mass of each person cloud and, given the angular symmetry of the human body, we used cylindrical coordinates. Finally, the person cloud was quantised into 3D bins, where each bin accumulates the mean color of the nearby data. With this particular representation, it was easy to accumulate all the different pointclouds of a person over time to provide a low-level person representation vector. Although this prior representation was a very rich feature vector, it presented a high dimensionality of data that was difficult to manage, and also presented strong correlation and noise. To avoid these problems and to make this vector more handy, we proposed different marginalisation techniques that compress the information to provide more robust and compact features. The marginalisation was carried out over the radius, angle, height, time or color. Depending on the marginalised variable, we got different descriptors. In total, we introduced four different methods.

The first method was the Bodyprint. These features generalise the person appearance by marginalysing time, radius and angle to represent the person appearance into a single height vector with rgb features. With this compact feature vector, we carried out preliminary evaluation, such as the influence of the color space or the use of contextual cues, which could be also extended to the other family of descriptors.

Then we introduced the 3D Bodyprints, which basically are Bodyprints that distribute the color features in height and angle. We demonstrated that these features were specially indicated when the person appearance was not symmetric in angle, so distributing the features in angles helped in the matching process by just comparing the features at the same angles. However, its use was not indicated in complex scenarios where the person trajectory and center of mass could not be estimated correctly, which produced a phase shifting that strongly affected the matching.

The following descriptor that we presented was the BoA - Bag of Appearances method. This method is a container of color features that collects all the different person appearances over time. This descriptor lets capture all the different appearances of a person along its tracking. The method showed special robustness against

outliers, which would strongly contaminate the Bodyprints while averaging all the temporal information into a single vector.

Finally, we introduced the latent models, which consist of a set of probabilistic techniques for dimensionality reduction that minimise the features correlation and noise, to provide a more compact feature vector, named *latent features*. These techniques can be applied to Bodyprints, 3D Bodyprints and Boa. In this work, we evaluated these models using Bodyprints. The results showed that these features compact data, reduce the correlation and noise, and cope with missing data.

The comparison of all the family of descriptors against the state of the art over our particular databases demonstrated that the proposed methods outperform any of the compared methods. In general, there are several reasons that make this family of methods robust for challenging conditions. First, the high precision of depth measurements allow placing each color pixel in the right position in space, which permits an accurate segmentation and intrinsically cope with momentary partial occlusions. Second, all the proposed methods take the temporal information of the person as input, which provides complete information of the person appearance when processed correctly. Third, the proposed illumination compensation module provides stability. Last, the use of the weighting vector that provides information of the reliability of the features has improved the results.

## 5.2 Future Work

In this work we have addressed the problem of people re-identification using rgb-depth cameras. We presented a full solution to the problem, covering image acquisition, segmentation, tracking, description and matching, focusing on the assessment of the proposed descriptor methods, which is the most challenging part. Since the use of depth cameras for commercial applications is quite new, there is yet too much work to be done in this area.

One of the current problems that is common to all of the descriptors is that we are assuming a Gaussian distribution of the color features at each vector component. However, this is a strong assumption, specially in the case of Bodyprints, BoA and latent features. In these descriptors, the mean color is calculated at each horizontal stripe for each height. However, in most of the stripes the color distribution will be multimodal, since the clothes can mix with the skin at a single stripe. For this reason, our future work will focus on the use of more complex appearance models that can cope with this multi-modality.

Using color features for discriminating people is a good approach when people wear different and distinctive cloths. However, there are a great number of people wearing clothes with similar colors. It is quite normal to find people wearing jeans and black coat, for example. Although the information of the person

height is considered in the matching process since it provides useful information, sometimes it is not enough distinctive. In these cases, other information such as gait analysis could help in the decision, since the gait is a distinctive biometrical cue. On other hand, features based on histograms of gradient orientations such as SIFT or SURF could be used to recognise interest points (such as logos on T-shirts). As the distance and orientation to the camera is known, the scale and rotation invariance could be disabled in these descriptors to be more distinctive.

The proposed methods have considered the temporal, angular and vertical distribution of the color features to create robust appearance descriptors. However, none of them used the radial information in the description, even though this information may be distinctive enough. Therefore, further work should be done considering the radial information. For example, one valid approach would be calculating the LBP on the radial representation of the person to catch the relative changes in the person shape.

Appendix A

# Dataset description

## A.1 General description

This appendix describes the two personal databases created in this Thesis. The reason that motivates the compilation of these databases arises with the fact that no other public databases containing rgb+depth information are found in the literature for the purpose of people re-identification.

The databases presented in this work collect people behaving normally in two different indoor scenarios: a school hall and a supermarket. The first database consists of a simpler case that contains a set of people walking straight ahead in a controlled scenario. The second database contains a more challenging scenario where people move randomly and may interact with objects in the scene.

The camera network topology used in these scenarios consist of two kinect cameras that cover different non-overlapped areas. Cameras capture different perspectives and have different illumination conditions. Cameras provide an aligned and synchronised pairwise of rgb+depth images at $320x240$ pixels of resolution. The frame rate varies with the scenario depending on the speed of the computer cpu and the type of its disk, which is a limiting factor to store images at high speed. For the school scenario, the frame rate was of 15fps. For the supermarket, only 6fps were possible.

In the database, Depth, RGB and mask images are provided for each camera node in the network, where the mask images represent the groundtruth of the dataset. There is one mask per frame, and each mask contains all existing person silhouettes at current frame. Silhouettes are filled with a value that represents its local ID (See Figure A.1). There is a unique local ID representing a single person over time for a particular camera node. This person may have a different local ID in the other nodes of the network. In order to be able to match the local ID's of the people in the database, a file with the ID's correspondences is provided with the database.

On other hand, the calibration information for each camera in the network is provided. This information contains the Rotation and Translation matrix of the camera regarding the world coordinate center, which is placed onto the floor with $z$ axis pointing perpendicular to it. The Translation matrix is just a 1D-vector containing the height of the camera regarding to the floor plane.

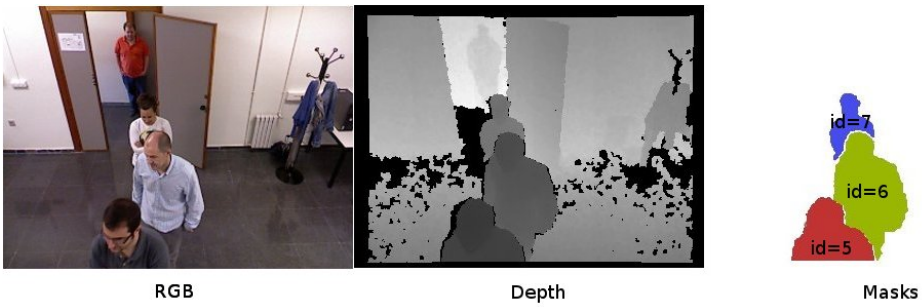



Figure A.1: Example of the rgb, depth and groundtruth images for a particular frame.

## A.2 Image Formats and Representation

Each different image that represent a frame in the sequence is stored using a prefix followed by the frame number. Rgb images are named using the prefix *Img*; depth images use the prefix *Depth*, and Mask images use *Mask* prefix.

Each of these images have different format. Rgb images are JPEG files. Mask images are uncompressed TIFF files. Depth images are presented using CIMG format, which can store floating images while compressing the information without loss. This formatting has been chosen in order speed up the process of data saving to disk, which is the bottle neck for having a high frame rate in the capture process. In these images, a value of zero in one pixel means that there is not available depth information for that coordinate.

CIMG format can be read using cimg Library [CIMG ] or the function cimgread.m, which is provided with the database, using MATLAB. The format of the cimg file contains two lines of text that contain the number of rows and columns. Using two-byte unsigned short numbers, the data matrix ordered by rows (top row first).

On other hand, rgb and depth pairwise images can be represented together as a pointcloud. PCL [Pointcloud ] library offers a wide range of functions for 3D data processing and visualization. In this Thesis, only visualization modules have been used for representing the clouds.

## A.3 School Hall Database

The school hall database contains a collection of images gathered in the Telecommunication building of the Technical University of Valencia. This dataset contains a simple scenario where people just walk through the cameras in straight directions.

Two cameras have been placed on the top of a door to cover the both sides of the door, with non-overlapping area. One camera covers the entrance and the other the exit, as can be seen in Figure A.2. People pass under the camera each time. Each camera has different position and orientation, and illumination may vary between

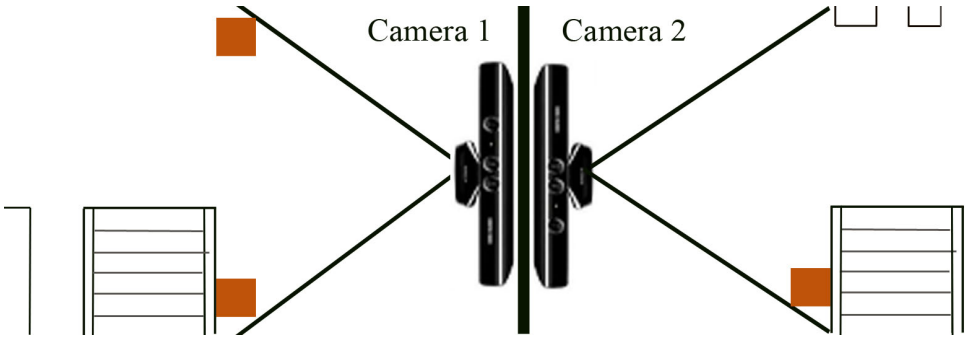



Figure A.2: Plan of the system deployment in the school hall.

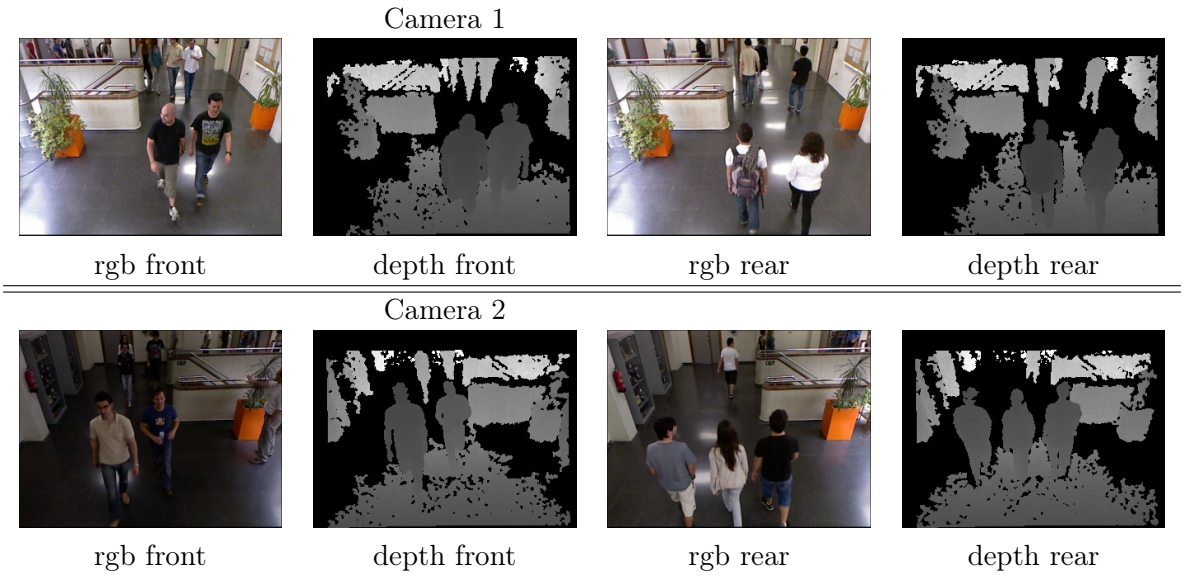



Figure A.3: Example of rgb and depth images taken from the two camera nodes in the school scenario containing front and rear views of people.

them but also over time due to the auto-gain property of kinect cameras.

The database is organised as follows: there is a set of images for training and another for testing. Both training and test images have the same people, but they appear with different poses. The training set contains 43 different people, which have been collected from frontal and rear views taken at different instants, as shown in Figure A.3. The test dataset contains the same 43 people but taken at other different instants.

## A.4 Supermarket Database

The supermarket database compiles a set of images gathered in Maxicarne supermarket in Paterna, Valencia. In the supermarket there are two cameras, one placed at the entrance and other at the exit, as seen in Figure A.4. At entrance, the camera

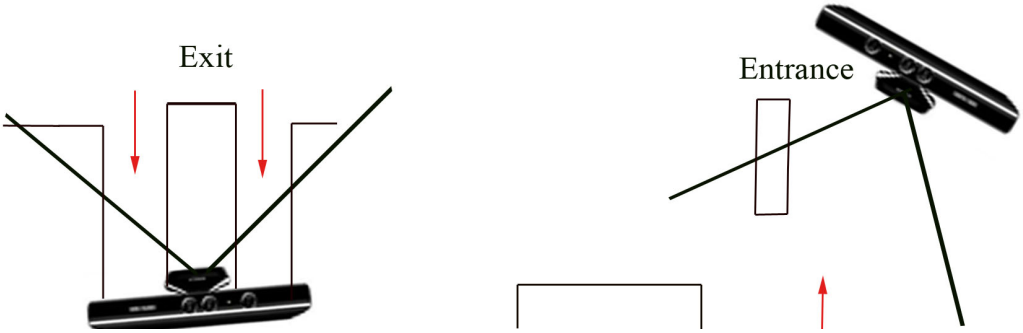



Figure A.4: Plan of the system deployment in the supermarket.

captures the left side of the clients from a tilt view. At exit, the camera covers two checkout tracks and covers frontal appearance.

The database is organised as follows: In the training dataset, 63 different people have been collected getting into the supermarket through the camera at the entrance and leaving through the camera at the checkout lanes. In the testing dataset, 73 different people have been collected in a different day than in the training set. People are also different between the datasets. In both cases, images have been taken during five hours.

We can say that this scenario is a challenging database for many reasons:

- People were not noticed about the system and they behaved freely, with no constrains about trajectories or movements
- Differences in the human behavior at the entrance and exit may exist due the use of the scene elements. In the entrance people may slightly bend the body to take a shopping basket from the pile, push a trolley taken from an outer part of the shop, walk fast, turn round or have non straight trajectories. On other hand, at the exit people can pass through the camera grabbing some items or pushing the trolley
- Lightning conditions are totally different for both cameras
- Cameras cover substantially different perspectives of the people

Figure A.5 shows some examples of people in the databased seen from the two cameras in the network. On other hand, Figure A.4.a and Figure A.4.b show more examples of people in the supermarket using pointcloud representation. The image on the left shows a woman holding a child at the entrance. The woman at the exit, who is seen from a different perspective, is pushing a trolley at the checkout lane.

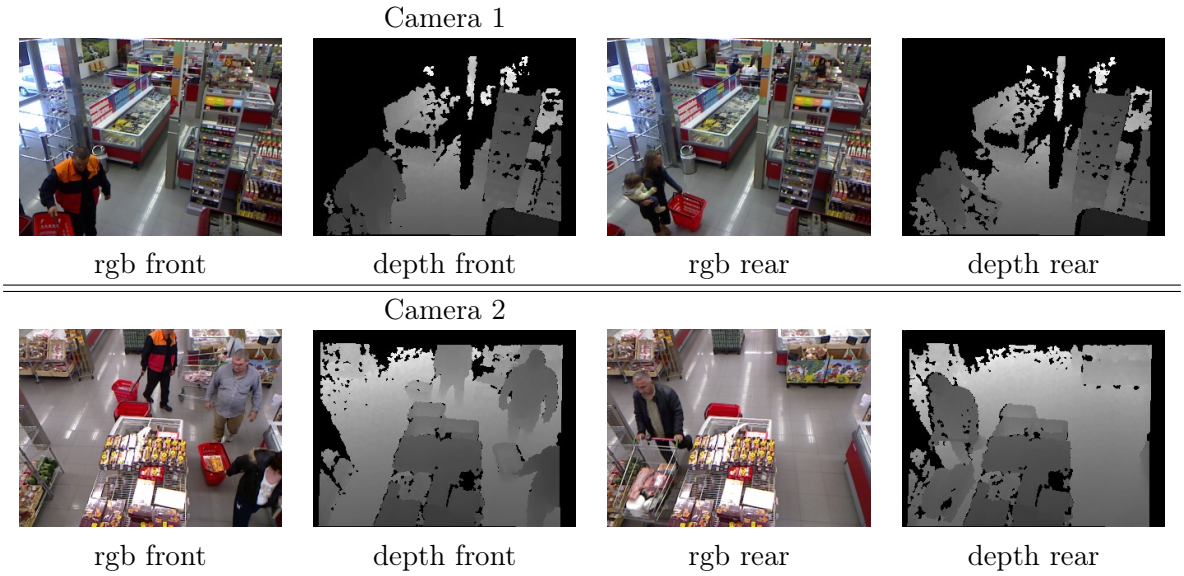



Figure A.5: Example of rgb and depth images taken from the two camera nodes in the school scenario containing front and rear views of people.

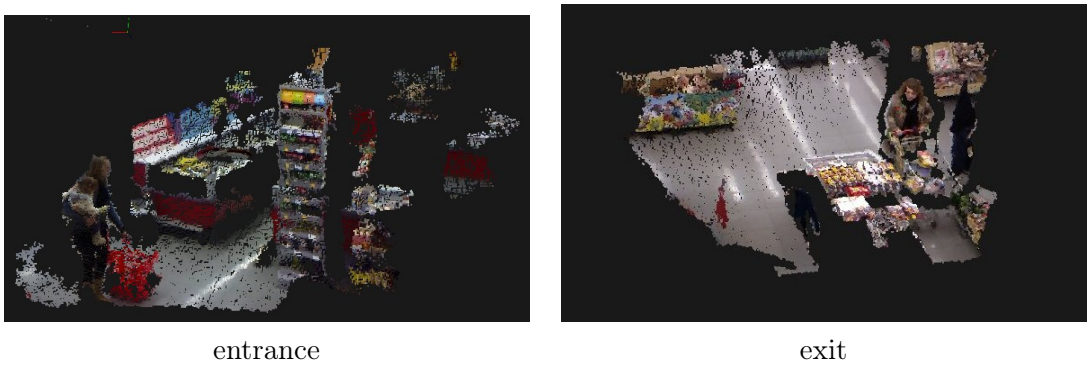



Figure A.6: Example of two rgb-depth image pairs represented using pointcloud library for the entrance and exit cameras.

Appendix B

# Matching Metrics

Depending on the nature of the data, different similarity measures may provide different results. In this Thesis the most representative matching metrics have been considered. The following lines describe the used metrics. Note that the weighted version of the similarity measures have also been included.

Let us define two feature vectors belonging to a probe and gallery set, respectively represented as $x_i = [x_i^1, ..., x_i^k, ..., x_i^N]$ and $x_j = [x_j^1, ..., x_j^k, ..., x_j^N]$. Let $w_i = [w_i^1, ..., w_i^k, ..., w_i^N]$ and $w_j = [w_j^1, ..., w_j^k, ..., w_j^N]$ be the corresponding weights. Let us define the following terms:

- $\mu_i = \frac{1}{N}\sum_k x_i^k$ as the mean value of $x_i$
- $\sigma_i = \frac{1}{N}\sum_k (x_i^k - \mu_i)^2$ as the variance of $x_i$
- $\mu_i' = \frac{\sum_k w_i^k x_i^k}{\sum_k w_i^k}$ as the weighted mean of $x_i$
- $\sigma_i' = \frac{\sum_k w_i^k (x_i^k - \mu_i)^2}{\sum_k w_i^k}$ as the weighted variance of $x_i$

The following matching metrics are defined:

- Correlation: A measure of the interdependence of two random variables that ranges in value from -1 to +1, indicating perfect negative correlation at -1, absence of correlation at zero, and perfect positive correlation at +1.

$$r_{corr}(x_i; x_j) = \frac{\frac{1}{n}\sum_k x_i^k x_j^k - \mu_i \mu_j}{\sigma_i \sigma_j} \tag{B.1}$$

- Weighted correlation: A measure of the interdependence of two random variables that ranges in value from -1 to +1, indicating perfect negative correlation at -1, absence of correlation at zero, and perfect positive correlation at +1.

$$r'_{corr}(w_i, x_i; w_j, x_j) = \frac{\frac{1}{\sum_k w_i^k w_j^k}\sum_k w_i^k w_j^k x_i^k x_j^k - \mu_i' \mu_j'}{\sigma_i' \sigma_j'} \tag{B.2}$$

- Euclidean distance: It is the basis of many measures of similarity and dissimilarity. Represents the shortest distance between two vectors. Distances range

from 0 to inf, where 0 indicates that the two vectors are the same; the higher the distance is, the more different the vectors are.

$$d_{Euc}(x_i; x_j) = \sqrt{\sum_{k=0}^{N}(x_i^k - x_j^k)^2} \tag{B.3}$$

- Weighted euclidean distance: It is the basis of many measures of similarity and dissimilarity. Represents the shortest distance between two vectors. Distances range from 0 to inf, where 0 indicates that the two vectors are the same; the higher the distance is, the more different the vectors are.

$$d'_{Euc}(x_i; x_j) = \sqrt{\frac{1}{\sum w_i^k w_j^k}\sum_{k=0}^{N}(w_i^k x_i^k - w_j^k x_j^k)^2} \tag{B.4}$$

- Cosine distance: Measures the cosine of the angle $\Theta_{ij}$ between two vectors. It measures the orientation and not the magnitude. Two vectors in the same orientation have cosine similarity 1; two vectors at 90º have similarity 0; two vectors totally opposed have similarity -1.

$$d_{cos}(x_i; x_j) = cos(\Theta_{ij}) = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|} = \frac{\sum_{k=0}^{N} x_i^k x_j^k}{\sqrt{\sum_{k=0}^{N}(x_i^k)^2}\sqrt{\sum_{k=0}^{N}(x_j^k)^2}} \tag{B.5}$$

- Weighted cosine distance: Measures the cosine of the angle $\Theta_{ij}$ between two vectors. It measures the orientation and not the magnitude. Two vectors in the same orientation have cosine similarity 1; two vectors at 90º have similarity 0; two vectors totally opposed have similarity -1.

$$d'_{cos}(x_i; x_j) = cos(\Theta_{ij}) = \frac{w_i x_i \cdot w_j x_j}{\|x_i\|\|x_j\|} = \frac{\sum_{k=0}^{N} w_i^k x_i^k w_j^k x_j^k}{\sqrt{\sum_{k=0}^{N}(w_i^k x_i^k)^2}\sqrt{\sum_{k=0}^{N}(w_j^k x_j^k)^2}} \tag{B.6}$$

- Mahalanobis distance: Measures the distance between two distribution of points, scaled by the statistical variation in each point component. It takes into account the covariance matrix $\sum$ of the sample distribution. This distance turns into Euclidean distance when the variance equals 1 in each of the axis.

$$d_{Mh}(x_i; x_j) = \delta_{ij} = \sqrt{\sum_{k=0}^{N}(x_i^k - x_j^k)^T \Sigma^{-1}(x_i^k - x_j^k)} \tag{B.7}$$

# Bibliography


[Alahi 2008] A. Alahi, D. Marimon, M. Bierlaire and M. Kunt. *A Master-Slave Approach for Object Detection and Matching with Fixed and Mobile Cameras*. In 15th IEEE International Conference on Image Processing, San Diego, 2008. (Cited on pages 20 and 21.)

[Alahi 2010] Alexandre Alahi, Pierre Vandergheynst, Michel Bierlaire and Murat Kunt. *Cascade of descriptors to detect and track objects across any network of cameras*. Comput. Vis. Image Underst., vol. 114, no. 6, pages 624–640, June 2010. (Cited on page 25.)

[Anguelov 2005] Dragomir Anguelov, Ben Taskar, Vassil Chatalbashev, Daphne Koller, Dinkar Gupta, Geremy Heitz and Andrew Ng. *Discriminative Learning of Markov Random Fields for Segmentation of 3D Scan Data*. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, CVPR '05, pages 169–176, Washington, DC, USA, 2005. IEEE Computer Society. (Cited on page 18.)

[Ankerst 1999] Mihael Ankerst, Gabi Kastenmuller, Hans-Peter Kriegel and Thomas Seidl. *3D shape histograms for similarity search and classification in spatial databases*. In SSD'99, pages 207–226. Springer, 1999. (Cited on page 23.)

[Baeuml 2011] Martin Baeuml and Rainer Stiefelhagen. *Evaluation of Local Features for Person Re-Identification in Image Sequences*. In IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), page 6, Aug. 2011. (Cited on page 22.)

[Bak 2010a] S. Bak, E. Corvee, F. Bremond and M. Thonnat. *Person Re-identification Using Haar-based and DCD-based Signature*. In 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, AMMCSS 2010, in conjunction with 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. AVSS, 2010. (Cited on page 22.)

[Bak 2010b] S. Bak, E. Corvee, F. Bremond and M. Thonnat. *Person Re-identification Using Spatial Covariance Regions of Human Body Parts*. In Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance, pages 435–440, september 2010. (Cited on page 22.)

[Bak 2011] S. Bak, E. Corvee, F. Bremond and M. Thonnat. *Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid*. In Advanced Video and Signal-Based Surveillance, Klagenfurt, Autriche, August 2011. (Cited on page 21.)

[Baltieri 2011] D. Baltieri, R. Vezzani, R. Cucchiara, A. Utasi, C. BenedeK and T. Szirányi. *Multi-view people surveillance using 3D information.* In ICCV Workshops, pages 1817–1824, 2011. (Cited on pages 16 and 24.)

[Barbosa 2012] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani and V. Murino. *Re-identification with RGB-D sensors.* In First International Workshop on Re-Identification, October 2012. (Cited on page 105.)

[Basilevsky 1994] A. Basilevsky. *Statistical factor analysis and related methods: theory and applications.* 1994. (Cited on page 25.)

[Bäuml 2010] M. Bäuml, K. Bernardin, K. Fischer, H.K. Ekenel and R. Stiefelhagen. *Multi-Pose Face Recognition for Person Retrieval in Camera Networks.* In International Conference on Advanced Video and Signal-Based Surveillance, 2010. (Cited on page 19.)

[Bazzani 2010] L. Bazzani, M. Cristani, A. Perina, M. Farenzena and V. Murino. *Multiple-Shot Person Re-identification by HPE Signature.* In Proceedings of the 2010 20th International Conference on Pattern Recognition, pages 1413–1416, Washington, DC, USA, 2010. (Cited on pages 16, 20 and 21.)

[Bedagkar-Gala 2011] Apurva Bedagkar-Gala and Shishir K. Shah. *Multiple person re-identification using part based spatio-temporal color appearance model.* In Computational Methods for the Innovative Design of Electrical Devices'11, pages 1721–1728, 2011. (Cited on page 26.)

[Benezeth 2008] Yannick Benezeth, Pierre-Marc Jodoin, Bruno Emile, Helene Laurent and Christophe Rosenberger. *Review and evaluation of commonly-implemented background subtraction algorithms.* In ICPR, pages 1–4. IEEE, 2008. (Cited on page 17.)

[Beucher 1992] S. Beucher and F. Meyer. The morphological approach to segmentation: The watershed transformation, pages 433–481. Marcel-Dekker, 1992. (Cited on page 39.)

[Bird 2005] N.D. Bird, O. Masoud, N.P. Papanikolopoulos and A. Isaacs. *Detection of loitering individuals in public transportation areas.* IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 2, pages 167–177, june 2005. (Cited on pages 21, 57 and 59.)

[Bishop 2006] C.M. Bishop. Pattern recognition and machine learning (information science and statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. (Cited on pages 94 and 96.)

[Bouchrika 2009] Imed Bouchrika, John N. Carter and Mark S. Nixon. *Recognizing People in Non-Intersecting Camera Views.* In International Conference on Imaging for Crime Detection and Prevention, 2009. (Cited on page 19.)

[Boureau 2010] YLan Boureau, Jean Ponce and Yann Lecun. *A Theoretical Analysis of Feature Pooling in Visual Recognition.* 2010. (Cited on page 45.)

[Bouwmans 2008] T. Bouwmans, F. El Baf, and B. Vachon. *Background modeling using mixture of gaussians for foreground detection - a survey.* In Recent Patents on Computer Science, 2008. (Cited on page 16.)

[Burges 1998] J.C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition.* In Data Min. Knowl. Discov., volume 2, pages 121–167, Hingham, MA, USA, June 1998. Kluwer Academic Publishers. (Cited on page 17.)

[Bustos 2005] B. Bustos, D.A. Keim, D. Saupe, T. Schreck and D.V. Vranic. *Feature-based similarity search in 3D object databases.* ACM Computing Surveys, vol. 37, page 2005, 2005. (Cited on page 23.)

[Cai 2008] Yinghao Cai, Kaiqi Huang and Tieniu Tan. *Human appearance matching across multiple non-overlapping cameras.* In ICPR, pages 1–4, 2008. (Cited on page 22.)

[Chen 2007] D. Chen, A. Bharucha and H. Wactlar. *People Identification through Ambient Camera Networks.* In International Conference on Multimedia and Ambient Intelligence, 2007. (Cited on page 19.)

[Cheng 2009] Y.M. Cheng, W.T. Zhou, Y. Wang, C.H. Zhao and S.W. Zhang. *Multi-Camera-Based Object Handoff Using Decision-Level Fusion.* In CISP09, pages 1–5, 2009. (Cited on page 20.)

[Chikahito 2003] Nakajima Chikahito, Pontil Massimiliano, Heisele Bernd and Poggio Tomaso. *Full-body person recognition system.* Pattern Recognition, vol. 36, no. 9, pages 1997–2006, 2003. (Cited on page 19.)

[CIMG ] CIMG. *CIMG library.* http://cimg.sourceforge.net/. (Cited on page 114.)

[Comaniciu 2003] D. Comaniciu, V. Ramesh and P. Meer. *Kernel-Based Object Tracking.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 5, pages 564–575, May 2003. (Cited on page 25.)

[Dalal 2005] N. Dalal and B. Triggs. *Histograms of Oriented Gradients for Human Detection.* In CVPR (1), pages 886–893, 2005. (Cited on pages 17, 22 and 28.)

[Dikmen 2010] M. Dikmen, E. Akbas, T.S. Huang and N. Ahuja. *Pedestrian recognition with a learned metric.* In Asian Conference in Computer Vision, 2010. (Cited on pages 21 and 26.)

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

[Duda 2007] R.O. Duda, P.E. Hart and D.G. Stork. *Extended Gaussian Images*. Pattern Classification, New York, vol. 24, no. 2, pages 305–307, September 2007. (Cited on page 23.)

[Elgammal 2000] Ahmed Elgammal, David Harwood and Larry Davis. *Non-parametric model for background subtraction*. In FRAME-RATE WORKSHOP, IEEE, pages 751–767, 2000. (Cited on page 17.)

[Farenzena 2010] 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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, CA, USA, 2010. IEEE Computer Society. (Cited on pages 21, 22, 26, 105 and 106.)

[Fei-Fei 2005] L. Fei-Fei and P. Perona. *A Bayesian hierarchical model for learning natural scene categories*. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 524 – 531 vol. 2, june 2005. (Cited on page 81.)

[Fischler 1981] Martin A. Fischler and Robert C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Commun. ACM, vol. 24, no. 6, pages 381–395, June 1981. (Cited on page 34.)

[Fisher 1936] R. A. Fisher. *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, vol. 7, no. 7, pages 179–188, 1936. (Cited on page 84.)

[Fodor 2002] I. Fodor. *A Survey of Dimension Reduction Techniques*. Rapport technique, 2002. (Cited on page 24.)

[Forssen 2007] Per-Erik Forssen. *Maximally Stable Colour Regions for Recognition and Matching*. In IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, June 2007. IEEE Computer Society, IEEE. (Cited on page 22.)

[Gandhi T. 2006] Gandhi T. and Trivedi M. *Panoramic Appearance Map (PAM) for Multi-camera Based Person Re-identification*. In International Conference on Video and Signal Based Surveillance, 2006. (Cited on page 16.)

[Gandhi 2007] Tarak Gandhi and Mohan Trivedi. *Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation*. Mach. Vision Appl., vol. 18, no. 3, pages 207–220, May 2007. (Cited on page 24.)

[Gerrard n 42] G. Gerrard and R. Thompson. How many cameras are in the uk. CCTV Image, 2011, n 42. (Cited on page 1.)

[Gheissari 2006] N. Gheissari, T.B. Sebastian and R. Hartley. *Person Reidentification Using Spatiotemporal Appearance*. In CVPR (2), pages 1528–1535, 2006. (Cited on pages 19, 21 and 25.)

[Gibson 9598] J.J. Gibson. The ecological approach to visual perception. 1986 , ISBN-10: 0898599598. (Cited on page 3.)

[Gilbert 2006] A. Gilbert and R. Bowden. *Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity*. In Proceedings of the 9th European conference on Computer Vision - Volume Part II, ECCV'06, pages 125–136, Berlin, Heidelberg, 2006. Springer-Verlag. (Cited on pages 20 and 21.)

[Gray 2007] D. Gray, S. Brennan and H. Tao. *Evaluating Appearance Models for Recognition, Reacquisition, and Tracking*. In Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), 2007. (Cited on page 55.)

[Gray 2008] D. Gray and H. Tao. *Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features*. In Proceedings of the 10th European Conference on Computer Vision: Part I, pages 262–275, Berlin, Heidelberg, 2008. (Cited on page 21.)

[Green 1999] M. W. Green. *The appropriate and effective use of security technologies in u.s. schools*. Rapport technique, Technical Report ncj 178265, Sandia National Laboratories, 1999. (Cited on page 1.)

[Grimaud 1992] M. Grimaud. *A new measure of contrast: the dynamics*. In SPIE, editeur, Image Algebra and Morphological Image Processing III, pages 292–305, 1992. (Cited on page 39.)

[Gulshan 2011] V. Gulshan, V. Lempitsky and A. Zisserman. *Humanising GrabCut: Learning to segment humans using the Kinect*. In IEEE Workshop on Consumer Depth Cameras for Computer Vision ICCV, 2011. (Cited on page 18.)

[Hahnel 2004] M. Hahnel, D. Klunder and K.F. Kraiss. *Color and texture features for person recognition.* In EEE International Joint Conference on Neural Networks, 2004. (Cited on page 20.)

[Ham 2008] Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, 2008. (Cited on page 22.)

[Hansen 2007] W. Hansen. *Automatic Detection of Zenith Direction in 3D Point Clouds of Built-Up Areas*. In PIA07 - Photogrammetric Image Analysis, pages 93–97, 2007. (Cited on page 32.)

[Haritaoglu 2000] I. Haritaoglu, D. Harwood and L. Davis. *W4-real time detection and tracking of people and their parts.* In IEEE Trans. PAMI, 2000. (Cited on page 17.)

[Hernandez-Vela 2012] A. Hernandez-Vela, M. Reyes, V. Ponce and S. Escalera. *GrabCut-Based Human Segmentation in Video Sequences.* Sensors, vol. 12, no. 11, pages 15376–15393, 2012. (Cited on page 17.)

[Hilaga 2001] M. Hilaga, Y. Shinagawa, T. Kohmura and T. L. Kunii. *Topology matching for fully automatic similarity estimation of 3D shapes.* In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, SIGGRAPH '01, pages 203–212, New York, NY, USA, 2001. ACM. (Cited on page 23.)

[Hirzer 2012] M. Hirzer, P.M. Roth and H. Bischof. *Person Re-identification by Efficient Impostor-Based Metric Learning.* In Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on, pages 203 – 208, 2012. (Cited on pages 25 and 26.)

[hsun Chang 2001] Ting hsun Chang and Shaogang Gong. *Tracking Multiple People with a Multi-Camera System.* In IEEE Workshop on Multi-Object Tracking, 2001. (Cited on page 21.)

[Hu 2006] Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou and S. Maybank. *Principal axis-based correspondence between multiple cameras for people tracking.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 4, pages 663 –671, april 2006. (Cited on page 25.)

[Ilin 2010] A. Ilin and T. Raiko. *Practical Approaches to Principal Component Analysis in the Presence of Missing Values.* J. Mach. Learn. Res., vol. 99, pages 1957–2000, August 2010. (Cited on page 97.)

[Jackson 2008] Jeremy D. Jackson, Anthony J. Yezzi and Stefano Soatto. *Dynamic Shape and Appearance Modeling via Moving and Deforming Layers.* Int. J. Comput. Vision, vol. 79, no. 1, pages 71–84, August 2008. (Cited on page 19.)

[Javed 2003] Omar Javed, Zeeshan Rasheed, Khurram Shafique and Mubarak Shah. *Tracking Across Multiple Cameras With Disjoint Views.* In Ninth IEEE International Conference on Computer Vision, pages 952–957, 2003. (Cited on pages 19 and 25.)

[Javed 2005] Omar Javed. *Appearance Modeling for Tracking in Multiple Non-overlapping Cameras.* In In IEEE International Conference on Computer Vision and Pattern Recognition, pages 26–33, 2005. (Cited on page 19.)

[Javed 2008] O. Javed, O. Shafique, Z. Rasheed and M. Shah. *Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping*

*views*. Comput. Vis. Image Underst., vol. 109, no. 2, pages 146–162, February 2008. (Cited on pages 20 and 21.)

[Jeong 2008] K. Jeong and C. Jaynes. *Object matching in disjoint cameras using a color transfer approach*. Mach. Vision Appl., vol. 19, no. 5-6, pages 443–455, September 2008. (Cited on page 20.)

[Jungling K. 2011] Jungling K., Bodensteiner, C. and Arens, M. *Person re-identification in multi-camera networks*. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, pages 55 –61, june 2011. (Cited on page 22.)

[Kang 1993] Sing Bing Kang and Katsushi Ikeuchi. *The Complex EGI: A New Representation for 3-D Pose Determination*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 7, pages 707–721, July 1993. (Cited on page 23.)

[Kang 2004] Jinman Kang, Isaac Cohen and Gerard Medioni. *Object Reacquisition Using Invariant Appearance Model*. In Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4 - Volume 04, ICPR '04, pages 759–762, Washington, DC, USA, 2004. IEEE Computer Society. (Cited on page 25.)

[Kettnaker 1999] Vera Kettnaker and Ramin Zabih. *Bayesian Multi-Camera Surveillance*. In 1999 Conference on Computer Vision and Pattern Recognition (CVPR 99), 23-25 June 1999, Ft. Collins, CO, USA, page 2253. IEEE Computer Society, 1999. (Cited on page 21.)

[Khoshelham 2012] Kourosh Khoshelham and Sander Oude Elberink. *Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications*. Sensors, vol. 12, no. 2, page 1437, 2012. (Cited on page 74.)

[Kohavi 1995] Ron Kohavi. *A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection*. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. (Cited on page 53.)

[Kortgen 2003] M. Kortgen, G. J. Park, M. Novotni and R. Klein. *3D Shape Matching with 3D Shape Contexts*. In The 7th Central European Seminar on Computer Graphics, April 2003. (Cited on page 23.)

[Krystian 2005] Mikolajczyk Krystian and Schmid Cordelia. *A Performance Evaluation of Local Descriptors*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 10, pages 1615–1630, October 2005. (Cited on page 19.)

[Kuo 2010] C.H. Kuo, C Huang and R. Nevatia. *Inter-camera association of multi-target tracks by on-line learned appearance affinity models*. In Proceedings of

the 11th European conference on Computer vision: Part I, ECCV'10, pages 383–396, Berlin, Heidelberg, 2010. Springer-Verlag. (Cited on page 21.)

[Kviatkovsky 2012] Igor Kviatkovsky. *Master Thesis. Color Invariants for Person Re-Identification*, 2012. (Cited on page 60.)

[Lan 2014] Rushi Lan, Yicong Zhou, Yuan Yan Tang and C.L.P. Chen. *Person reidentification using quaternionic local binary pattern.* In Multimedia and Expo (ICME), 2014 IEEE International Conference on, pages 1–6, July 2014. (Cited on page 22.)

[Liang 1999] Z. Liang and T. Chuck. *Stereo- and Neural Network-Based Pedestrian Detection.* IEEE Trans. on Intelligent Transportation Systems, vol. 1, pages 148–154, 1999. (Cited on page 18.)

[Lyer 2005] Natraj Lyer, Subramaniam Jayanti, Kuiyang Lou, Yagnanarayanan Kalyanaraman and Karthik Ramani. *Three-dimensional shape searching: state-of-the-art review and future trends.* Comput. Aided Des., vol. 37, no. 5, pages 509–530, April 2005. (Cited on page 23.)

[Ma 2012] Bingpeng Ma, Yu Su and Frédéric Jurie. *Local Descriptors Encoded by Fisher Vectors for Person Re-identification.* In ECCV Workshops (1), pages 413–422, 2012. (Cited on page 25.)

[Madden 2007] C. Madden, E. Cheng and M. Piccardi. *Tracking people across disjoint camera views by an illumination-tolerant appearance representation.* Machine Vision and Applications, vol. 18, pages 233–247, 2007. (Cited on pages 21 and 26.)

[Marr 67 8] D. Marr. A computational investigation into the human representation and processing of visual information. ISBN 0-7167-1567-8. (Cited on page 3.)

[Martinel 2012] N. Martinel and C. Micheloni. *Re-identify people in wide area camera network.* In IEEE, editeur, Computer Vision and Pattern Recognition Workshops (CVPRW), pages 31–36, Providence RI, June 2012. (Cited on page 19.)

[Mazzon 2012] R. Mazzon, S.F. Tahir and A. Cavallaro. *Person re-identification in crowd.* Pattern Recognition Letters, no. 0, pages –, 2012. (Cited on page 21.)

[Microsoft a] Microsoft. *Kinect inside.* http://gilotopia.blogspot.com.es/2010/11/how-does-kinect-really-work.html. (Cited on page 30.)

[Microsoft b] Microsoft. *Msoft.* http://kinectforwindows.org/. (Cited on pages 18 and 30.)

[Monari 2012] Eduardo Monari. *Color Constancy Using Shadow-Based Illumination Maps for Appearance-Based Person Re-identification.* In Advanced Video

and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on, pages 197 –202, sept. 2012. (Cited on page 21.)

[Novotni 2003] Marcin Novotni and Reinhard Klein. *3D zernike descriptors for content based shape retrieval*. In Proceedings of the eighth ACM symposium on Solid modeling and applications, SM '03, pages 216–225, New York, NY, USA, 2003. ACM. (Cited on page 23.)

[Oliveira 2009] I.O. Oliveira and J.L. Souza Pio. *People Reidentification in a Camera Network*. In Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pages 461–466, dec. 2009. (Cited on pages 21 and 22.)

[Oliver 2012a] Javier Oliver. *Database GPI*. http://www.gpiv.upv.es/kinect_data/, 2012. (Cited on page 40.)

[Oliver 2012b] Javier Oliver. *Database GPI with pointclouds*. http://www.gpiv.upv.es/kinect_data_cylinder/, 2012. (Cited on page 46.)

[Oltramari 2012] A. Oltramari and C. Lebiere. *Using Ontologies in a Cognitive-Grounded System: Automatic Action Recognition in Video Surveillance*. In Proceedings of the Seventh International Conference on Semantic Technology for Intelligence, Defense, and Security, 2012. (Cited on page 2.)

[Oncel 2006] T. Oncel, P. Fatih and M. Peter. *Region Covariance: A Fast Descriptor for Detection And Classification*. In In Proc. 9th European Conf. on Computer Vision, pages 589–600, 2006. (Cited on page 22.)

[OpenCV ] OpenCV. *OpenCV*. http://opencv.willowgarage.com/. (Cited on page 30.)

[Openni ] Openni. *Openni*. http://www.openni.org/. (Cited on page 30.)

[operators ] Cctv operators. *Cctv operators*. http://www.unitedcollege.org/cctv-course.html. (Cited on page 2.)

[Orwell 1999] J. Orwell, P. Remagnino and G.A. Jones. *Multi-camera colour tracking*. 2nd IEEE Workshop on Visual Surveillance, 1999. (Cited on page 20.)

[Papadakis 2010] P. Papadakis, I. Pratikakis, T. Theoharis and S.J. Perantonis. *PANORAMA: A 3D Shape Descriptor Based on Panoramic Views for Unsupervised 3D Object Retrieval*. International Journal of Computer Vision, vol. 89, no. 2-3, pages 177–192, 2010. (Cited on page 24.)

[Pedagadi 2013] Sateesh Pedagadi, James Orwell, Sergio A. Velastin and Boghos A. Boghossian. *Local Fisher Discriminant Analysis for Pedestrian Re-identification*. In CVPR, pages 3318–3325. IEEE, 2013. (Cited on page 25.)

[Pointcloud ] Pointcloud. *PCL ROS*. http://pointclouds.org/. (Cited on pages 30 and 114.)

[Porikli 2003] Fatih Murat Porikli. *Inter-camera color calibration by correlation model function*. In ICIP (2), pages 133–136, 2003. (Cited on page 21.)

[PrimeSense ] PrimeSense. *PrimeSense*. http://www.primesense.com/. (Cited on page 30.)

[Prosser B. 2010] Prosser B., Zheng W.S., Gong S. and Xiang T. *Person Re-Identification by Support Vector Ranking*. In Proceedings of the British Machine Vision Conference, pages 21.1–21.11. BMVA Press, 2010. (Cited on pages 20, 21, 24 and 26.)

[Prosser 2008] B. Prosser, S. Gong and T. Xiang. *Multi-camera matching under illumination change over time*. In in: European Conference on Computer Vision, 2008. (Cited on page 20.)

[ROS ] ROS. *ROS*. http://www.ros.org/. (Cited on page 30.)

[Roweis 1998] S. Roweis. *EM Algorithms for PCA and SPCA*. In in Advances in Neural Information Processing Systems, pages 626–632. MIT Press, 1998. (Cited on pages 25 and 95.)

[Salas 2011] Joaquin Salas and Carlo Tomasi. *People detection using color and depth images*. In Proceedings of the Third Mexican conference on Pattern recognition, MCPR'11, pages 127–135, Berlin, Heidelberg, 2011. Springer-Verlag. (Cited on pages 18 and 24.)

[Satta 2011] R. Satta, G. Fumera and F. Roli. *Exploiting Dissimilarity Representations for Person Re-Identification*. Venice, Italy, 28/09/2011 2011. (Cited on page 24.)

[Satta 2012] R. Satta, G. Fumera and F. Roli. *Fast person re-identification based on dissimilarity representations*. Pattern Recognition Letters, Special Issue on Novel Pattern Recognition-Based Methods for Reidentification in Biometric Context, vol. 33, pages 1838–1848, 10/2012 2012. (Cited on page 24.)

[Shotton 2011] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman and Andrew Blake. *Real-time human pose recognition in parts from single depth images*. In In In CVPR, 3, 2011. (Cited on page 18.)

[Stauffer 1999] Chris Stauffer and W. Eric L. Grimson. *Adaptive Background Mixture Models for Real-Time Tracking*. In CVPR, pages 2246–2252, 1999. (Cited on page 17.)

[Subhransu 2013] M. Subhransu, C.B. Alexander and M. Jitendra. *Efficient Classification for Additive Kernel SVMs*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pages 66–77, 2013. (Cited on page 17.)

[Sundar 2003] H. Sundar, D. Silver, N. Gagvani and S. Dickinson. *Skeleton Based Shape Matching and Retrieval*. In Proceedings of the Shape Modeling International 2003, SMI '03, pages 130–, Washington, DC, USA, 2003. IEEE Computer Society. (Cited on page 23.)

[Swain 1991] Michael J. Swain and Dana H. Ballard. *Color indexing*. International Journal of Computer Vision, vol. 7, pages 11–32, 1991. (Cited on page 21.)

[Tao 2015] Dapeng Tao, Lianwen Jin, Yongfei Wang and Xuelong Li. *Person Reidentification by Minimum Classification Error-Based KISS Metric Learning*. IEEE T. Cybernetics, vol. 45, no. 2, pages 242–252, 2015. (Cited on page 22.)

[Tapaswi 2012] Makarand Tapaswi, Martin Bäuml and Rainer Stiefelhagen. *Knock! Knock! Who is it? probabilistic person identification in TV-series*. In CVPR, pages 2658–2665, 2012. (Cited on page 16.)

[Tipping 1999] M.E. Tipping and C.M. Bishop. *Probabilistic Principal Component Analysis*. Journal of the Royal Statistical Society, Series B, vol. 61, pages 611–622, 1999. (Cited on page 25.)

[Tung 2005] Tony Tung and Francis Schmitt. *The Augmented Multiresolution Reeb Graph Approach for Content-based Retrieval of 3d Shapes*. International Journal of Shape Modeling, vol. 11, no. 1, pages 91–120, 2005. (Cited on page 23.)

[Vachier 1995] C. Vachier and F. Meyer. *Extinction value: a new measurement of persistence*. In IEEE Workshop on Nonlinear Signal and Image Processing, 1995. (Cited on page 39.)

[Vandergheynst 2009] P. Vandergheynst, M. Bierlaire, M. Kunt and A. Alahi. *Cascade of Descriptors to Detect and Track Objects Across Any Network of Cameras*. Computer Vision and Image Understanding, pages 1413–1416, 2009. (Cited on page 22.)

[Varma 2005] M. Varma and A. Zisserman. *A Statistical Approach to Texture Classification from Single Images*. International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis, vol. 62, no. 1–2, pages 61–81, April 2005. (Cited on page 81.)

[Verbeek 2009] Jakob Verbeek. *Notes on probabilistic PCA with missing values*. Rapport technique, Tech. report, 2009. (Cited on page 96.)

[Viola 2005] P. Viola, M.J. Jones and D. Snow. *Detecting Pedestrians Using Patterns of Motion and Appearance*. Int. J. Comput. Vision, vol. 63, no. 2, pages 153–161, July 2005. (Cited on page 17.)

[Wang 2006] H. Wang, D. Suter and K.Schindler. *Effective appearance model and similarity measure for particle filtering and visual tracking.* In Proceedings of the 9th European conference on Computer Vision - Volume Part III, ECCV'06, pages 606–618, Berlin, Heidelberg, 2006. Springer-Verlag. (Cited on page 20.)

[Wang 2007] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher and P. H. Tu. *Shape and appearance context modeling.* In iccv, pages 1–8, 2007. (Cited on pages 16 and 25.)

[Winn 2005] J. Winn, A. Criminisi and T. Minka. *Object categorization by learned universal visual dictionary.* In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 2, pages 1800 –1807 Vol. 2, oct. 2005. (Cited on page 81.)

[Wren 1997] Christopher Wren, Ali Azarbayejani, Trevor Darrell and Alex Pentland. *Pfinder: Real-Time Tracking of the Human Body.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pages 780–785, 1997. (Cited on page 17.)

[Xiao 2012] J. Xiao. *Contextual boost for pedestrian detection.* In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR '12, pages 2895–2902, Washington, DC, USA, 2012. IEEE Computer Society. (Cited on page 17.)

[Xu 2003] F. Xu and K. Fujimura. *Human Detection Using Depth and Gray Images.* In Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS '03, pages 115–, Washington, DC, USA, 2003. IEEE Computer Society. (Cited on page 18.)

[Yali 1996] Amit Yali and Kong Augustine. *Graphical Templates for Model Registration.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 3, pages 225–236, 1996. (Cited on page 19.)

[Yang 2011] Jun Yang, Zhongke Shi and Patricio A. Vela. *Person Reidentification by Kernel PCA Based Appearance Learning.* In Proceedings of the 2011 Canadian Conference on Computer and Robot Vision, CRV '11, pages 227–233, Washington, DC, USA, 2011. IEEE Computer Society. (Cited on page 24.)

[Yu 2007] Yang Yu, David Harwood, Kyongil Yoon and Larry S. Davis. *Human appearance modeling for matching across video sequences.* Mach. Vision Appl., vol. 18, no. 3, pages 139–149, May 2007. (Cited on page 25.)

[Zaharia 2001] T. Zaharia and F. Preteux. *Three-dimensional shape-based retrieval within the MPEG-7 framework.* In Proceedings SPIE Conference on Nonlinear Image Processing and Pattern Analysis XII, Vol. 4304, pp. 133-145, San Jose, CA, January, 2001. (Cited on page 23.)

[Zalevsky 2005] Zalevsky. *Kinect patent*, 2005. (Cited on page 30.)

[Zhang Z 2005] Zhang Z and Troje N.F. *View-independent person identification from human gait*. Neurocomputing, vol. 69, pages 250–256, 2005. (Cited on page 19.)

[Zhang 2007] L. Zhang, M.J. Fonseca and Ferreira. A survey on 3d shape descriptors. Technical Report FCT POSC/EIA/59938/2004, DecorAR, Lisboa Portugal, 2007. (Cited on pages 22 and 23.)

[Zhao 2005] T. Zhao, M. Aggarwal, R. Kumar and H. Sawhney. *Real-time wide area multi-camera stereo tracking*. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 976–983, 2005. (Cited on page 35.)

[Zheng 2011] W.S. Zheng, S. Gong and T. Xiang. *Person re-identification by probabilistic relative distance comparison*. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 649–656, june 2011. (Cited on pages 21, 26, 105 and 106.)

[Zhou 2001] Q. Zhou and J. Aggarwal. *Tracking and classifying moving objects from video*. In IEEE Proc. PETS Workshop, 2001. (Cited on page 17.)