Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari

Dissimilarity-based people re-identification and search for intelligent video surveillance

Riccardo Satta

Advisor: Prof. Fabio Roli
Co-advisor: Prof. Giorgio Fumera
Curriculum: ING-INF/05 - Sistemi di Elaborazione delle Informazioni

XXV Cycle
April 2013

Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari

Dissimilarity-based people re-identification and search for intelligent video surveillance

Riccardo Satta

Advisor: Prof. Fabio Roli
Co-advisor: Prof. Giorgio Fumera
Curriculum: ING-INF/05 - Sistemi di Elaborazione delle Informazioni

XXV Cycle
April 2013

To my family

Acknowledgments

“If opportunity doesn’t knock, build a door.”
Anonymous

Il Dottorato di Ricerca è una grande opportunità e insieme un gran rischio. Come il mio advisor (che ringrazierò opportunamente più avanti in queste righe) ama raccontare, è una “scatola vuota”: robusta, ampia, ma pur sempre all’inizio vuota. E senza le istruzioni per riempirla. Il successo del tuo Dottorato dipende allora da cosa ci metti dentro. I tuoi colleghi, il tuo advisor, il tuo gruppo di ricerca, saranno fondamentali per dare forma e peso al contenuto, ma – parliamoci chiaro – l’onore e l’onere di riempirla sta a te, e solo a te.

Guardandomi indietro, posso dire, questa scatola, di averla riempita per bene. Questi tre anni sono stati per me una grande, impagabile occasione di crescita professionale e personale. Per questo, come di rito, e doverosamente, desidero ringraziare chi mi ha accompagnato verso questo traguardo. In ordine sparso, inizio col dire grazie a Prof. Fabio Roli, che mi ha permesso di intraprendere questo percorso, mi ha incitato a dare sempre il massimo ed è stato una indispensabile guida nell’universo, non proprio facile, della ricerca. Ringrazio il paziente Prof. Giorgio Fumera, che ha supportato (e sopportato) per tre anni le mie “visioni” e mi ha aiutato a mantenerle entro i binari del possibile. Spero di aver carpito almeno un decimo della sua capacità analitica, sarebbe già sufficiente per una grande carriera nella ricerca. Desidero inoltre ringraziare Davide, Bat, Federico, Prof. Giorgio Giacinto, e i membri tutti del gruppo PRA. Tra questi un grazie particolare ad Ignazio, indispensabile “sparring partner”, per le frequenti discussioni e divagazioni che ci hanno permesso di partorire tante (forse troppe!) idee e progetti: prima o poi riusciremo a metterle tutte in pratica!

Oltre ai colleghi del gruppo PRA, con i quali ho condiviso gioie e dolori di questo percorso, ringrazio la mia famiglia e i miei amici, fondamentali per me in questi anni. Senza dubbio, mia madre Serena e mio padre Mariano meritano il ringraziamento più grande. È grazie a loro e al loro supporto se son potuto arrivare a questo traguardo. Ringrazio infine Patrizia, che mi è stata vicina questi tre anni, sopportando con santa pazienza i miei sfoghi, e ha sempre, e comunque, creduto in me.

A tutti, grazie.

Abstract

Intelligent video-surveillance is at present one of the most active research fields in computer science. It brings together a wide variety of computer vision and machine learning techniques to provide useful tools for surveillance operators and forensic video analytics. Person re-identification is among these tools; it consists of recognising whether an individual has already been observed over a network of cameras. Person re-identification has various possible applications, e.g., off-line retrieval of all the video-sequences showing an individual of interest whose image is given as query, or on-line pedestrian tracking over multiple cameras. The task is typically achieved by exploiting the clothing appearance, as classical biometric traits like the face are impractical in real-world video surveillance scenarios. Clothing appearance is represented by means of low-level local and global features of the images, usually extracted according to some part-based body model to treat different body parts (e.g. torso and legs) independently. The use of novel sensor technologies, e.g. RGB-D cameras like the MS Kinect, could also allow for the extraction of anthropometric measures from a reconstructed 3D model of the body, that can be used in combination with the clothing appearance to increase recognition accuracy.

This thesis presents a novel framework, named Multiple Component Dissimilarity (MCD), to construct descriptors of images of persons, using *dissimilarity representations*, a recent paradigm in machine learning in which the objects of interest are described as vectors of dissimilarities to a set of predefined prototypes. MCD extends the original dissimilarity paradigm to objects decomposable in multiple parts and with localised characteristics, to better deal with the peculiarities of the human body. The use of MCD has at least three important advantages:

(i) a drastic reduction of computational needs, mostly due to the compactness of dissimilarity representations (basically, small vectors of real numbers, easy to store and very fast to be matched);

(ii) a totally generic formulation of the underlying low-level representation, that allows one to combine different descriptors, even if they are heterogeneous in terms of the model and features used, into a single dissimilarity vector;

(iii) it provides a natural way to learn high-level concepts from low-level representations.

Building on its above salient features, MCD is used in this thesis to achieve several objectives:

(i) develop an approach to speed up existing person re-identification methods;

(ii) implement a novel person re-identification method based on the combination of different local and global features into a single dissimilarity vector, able to attain state-of-the-art performance;

(iv) develop a multi-modal approach to person re-identification (a novelty in the literature), by combining the clothing appearance with anthropometric measures extracted through the use of novel RGB-D sensors, into a single dissimilarity vector;

(v) develop a method to perform a novel task, proposed for the first time in this thesis, consisting in finding, among a set of images of individuals, those relevant to a *textual*, semantic query describing clothing appearance of an individual of interest. This task has been named *appearance-based people search* and can be useful in applications like forensics video analysis, where a textual description of the individual of interest given by a witness can be available, instead of an image.

Person re-identification and appearance-based people search are different tasks, aimed at addressing different problems. Still, they can be seen as instances of the more general problem of *searching and matching people on multi-media data*, e.g., video footages, range-depth data, speech audio data. Building on the commonalities with *Information Retrieval*, in the final part of the thesis, a possible formulation of the task of *people search on multi-media data* will be proposed, with some suggestions and guidelines on how to exploit the MCD framework for addressing this novel class of problems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The demand for security and safety of citizens and critical infrastructures is continuously growing in our society. Governments, international institutions and private companies are going along with these needs, spending a huge amount of efforts. A key role in this context is played by video-surveillance systems: nowadays, network of CCTV cameras have been deployed everywhere (Fig. 1.1). Camera networks are in principle an useful tool for addressing a variety of security issues [50]. E.g., for the prevention of crimes and of accidents (e.g. in an industrial facility), for the surveillance of state borders, for forensic investigations, and for the safeguard of the environment (e.g. forest fire detection). Last but not least, their presence can act as a strong deterrent for criminals. However, monitoring and analysing the massive quantity of recorded videos that a typical camera network generates per day, is becoming a critical problem. Surveillance operators are required to survey tens or hundreds of cameras at the same time; investigations that take place after a crime may need the review of hundreds of hours of footage.

To extract useful information from large collections of videos taken by surveillance cameras, and to help human operators in handling and understanding what is currently seen by a camera network, is the challenge of *intelligent video-surveillance*, which is at present one of the most active research fields in computer engineering and computer science. It brings together a wide variety of *computer vision* and *machine learning* techniques to enable various useful applications, such as:

- on-line tracking of the movements of a person or an object of interest [70, 100, 137];
- recognition of suspicious actions (e.g. a person running in the crowd) [109];
- detection of particular events of interest (e.g. a luggage being left unattended at an airport) [43, 131];
- summarisation of long footages to highlight only parts of potential interest [96].

Person re-identification[39] is another task that intelligent video-surveillance systems can enable. It consists of recognising an individual who has already been observed (hence the term *re*-identification) over a network of cameras. It is currently attracting much interest from researchers, due to its various possible applications, e.g., off-line retrieval of all the video-sequences where an individual of interest appears, whose image is given as query,

Figure 1.1: From left to right: a CCTV camera located at the Warwick Castle, an historical attraction in the country town of Warwickshire, UK; CCTV cameras in London, UK. According to a recent study [54], 1.85 million CCTV cameras operate in the sole UK.

or on-line pedestrian tracking over multiple, possibly not-overlapping cameras (a task also known as *re-acquisition* [57]).

While several biometric traits can be in principle used to this aim, strong pose variations and unconstrained environments (see Fig. 1.3) make the use of classical biometric traits like face difficult of impractical [39] with the typical sensors and setting of a surveillance network. Therefore, researchers explored the use of cues that pose less constraints, at the expense of an intrinsically lower identification capability. Among them, clothing appearance is used in the most of re-identification methods, as a soft, session-based cue, that is relatively easy to extract, and exhibits uniqueness over limited periods of time. Various *descriptors* of the clothing appearance have been proposed so far in the literature [39]. They are mostly designed heuristically, and are based on the extraction of various kinds of low-level *local* and *global* features from the images showing the individual[1]. Typically, they exploit a part-based body model, to take into account the non-rigid structure of the human body and treat the appearance of different body parts (e.g. torso and legs) independently.

Building on person re-identification, in this thesis another useful, novel task is proposed, that can be implemented using clothing appearance descriptors. The task has been named *appearance-based people search*, and consists in finding, among a set of images of individuals, the ones relevant to a *textual* query describing clothing appearance of an individual of interest. People search differs from person re-identification, as the query in this case is a textual, semantic description, instead of an image. This can be useful in applications like forensics video analysis, where a textual description of the individual of interest given by a witness can be available, instead of an image.

Apart from the clothing appearance, it is difficult to extract other cues from video streams of classical CCTV cameras. However, the extraction of other soft biometrics can be enabled by the recent introduction of combined video and range (RGB-D) sensors like MS Kinect [80]; for instance, they can be used to estimate various anthropometric measures useful to perform re-identification [10], like the height, the arm length, the leg length.

This thesis presents a novel framework to construct descriptors of the human appearance for the above tasks, using *dissimilarity representations* [104], a recent paradigm in machine learning in which the objects of interest are described as vectors of dissimilarities to a set of predefined *prototypes*. The framework, called Multiple Component Dissimilar-

[1] The term “local features” refers to localised characteristics of the image, e.g. the colour distribution around a certain salient point of the image; the term “global features”, instead, refers to characteristics of the whole image, e.g. the overall colour distribution.

ity (MCD), extends the original dissimilarity paradigm to objects decomposable in multiple parts and with localised characteristics, to better deal with the peculiarities of the human body. MCD allows for the construction of extremely compact representations, and carries at least three important advantages with respect to the tasks of person re-identification and people search:

- First, it can drastically reduce the issue of computational complexity, specially of the *matching* phase of person re-identification methods[2], due to the compactness of dissimilarity representations. It is worth noting that computational requirements have been almost overlooked so far; as a result, many re-identification methods are not suitable for direct deployment on real-world systems.
- Second, it builds upon a totally generic formulation of the underlying low-level representation, and therefore can be used to combine different descriptors, even if they are heterogeneous in terms of the model and features used. Such descriptors can also come from modalities different to the clothing appearance (e.g., the face, anthropometric measures obtained using a RGB-D sensors) Therefore, it can also be used to perform *multi-modal* person re-identification, in cases where the clothing appearance is not the only cue available.
- Third, it provides a natural way to learn high-level concepts from low-level representations. This directly enables the task of appearance-based people search described above.

Building on the above salient features of MCD, in this thesis the novel dissimilarity framework is exploited to:

- Develop a general approach to speed up any existing person re-identification method based on appearance body models with multiple body parts and/or local features, which includes most methods of the current literature.
- Implement a novel person re-identification method based on the combination, through MCD, of different local and global features, that attains state-of-the-art performance on common benchmark data sets.
- Combine the clothing appearance with anthropometric measures (e.g., the height, the arm length) extracted from the Depth information of RGB-D sensors, to perform *multi-modal* person re-identification. To the best of the Author's knowledge, the combined use of multiple soft modalities for person re-identification has never been proposed so far in the literature on this topic.
- Develop a general method to perform *appearance-based people search*, exploiting the same appearance descriptors used for person re-identification.

While person re-identification and people search can be addressed using the same MCD framework above, they are different tasks. However, they can be seen as instances of the

[2] *Matching* refers to the act of finding the most similar person in a gallery of candidates, given an image of the person of interest as query. The computational time required by the matching phase may become critical, especially in real-time applications, as the number of candidates grows.

more general problem of searching and matching people on multi-media data, i.e. video footages, range-depth data, but also speech audio data and so on. In the final part of the thesis, a possible formulation of the task of *people search on multi-media data* will be proposed, with some suggestions and guidelines on how to exploit the MCD framework for addressing this class of problems.

The rest of this introductory Chapter is structured as follows. First, in Sect. 1.1 a closer insight on person re-identification is given. Then, dissimilarity representations are introduced in Sect. 1.2. People search on multimedia data is discussed in Sect. 1.3. Finally, in Sect. 1.4 I underline the main contributions of this thesis and outline the rest of the work.

1.1 People re-identification and search

Knowing whether a person of interest was present in a given place at a given time is of crucial importance in many surveillance tasks. For this reason, researchers have spent a lot of efforts in developing techniques to detect people seen by a camera. To track an individual in a network of multiple cameras requires to maintain his/her identity over different fields of view (FOVs). Indeed, if FOVs are at least partly overlapping, one may exploit the simultaneous presence of a person in two or more video feeds to keep his/her identity over different cameras [79]. However, as the size of the site where the camera network is deployed grows, guaranteeing enough overlapping FOVs becomes difficult or even unsustainable.

Person re-identification is the task of associating video-sequences of people seen in different cameras, with generally non overlapping FOVs [39]. It is based on the extraction of signatures, or *descriptors*, associated to each tracked person in each camera view. When a tracked person leaves a certain camera view, and then reappears in a different one, the descriptors extracted from the latter camera view is matched against the former one, allowing one to reassign to that person the same identity that was previously associated to him/her. Re-identification is therefore born to associate *on-the-fly* different video-sequences to the same individual. Indeed, another possible application is to give a video-sequence, or an image, of an individual of interest as *query* to the system, and retrieve from a data base of previously stored video-sequences those showing the same individual. This task may be very useful e.g., for off-line investigations.

Formally, person re-identification is usually modelled as a classical *matching* problem, whose goal is to rank templates in a *template gallery* with respect to their similarity to a given *probe* individual. Thus, the problem of re-identifying an individual represented by its descriptor $\mathbf{P}$ can be formulated as:

$$\mathbf{T} = \arg\min_{\mathbf{T}_i} D(\mathbf{T}_i, \mathbf{Q}) \ , \mathbf{T}_i \in \mathscr{T} \qquad (1.1)$$

where $\mathscr{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_N\}$ is a gallery of N template descriptors, and $D(\cdot,\cdot)$ is a proper distance metric.

Descriptors are created from the sequence of rectangular regions (blobs) of frames containing the person, and can in principle be extracted using a variety of biometric cues. However, the typical re-identification setting is characterised by multiple, uncalibrated cameras in an unconstrained environment, with free poses and non collaborative users. This makes the use of classical biometric cues like face of gait not feasible in practice. For this reason, researchers have concentrated their efforts in exploiting the *clothing appearance* to con-

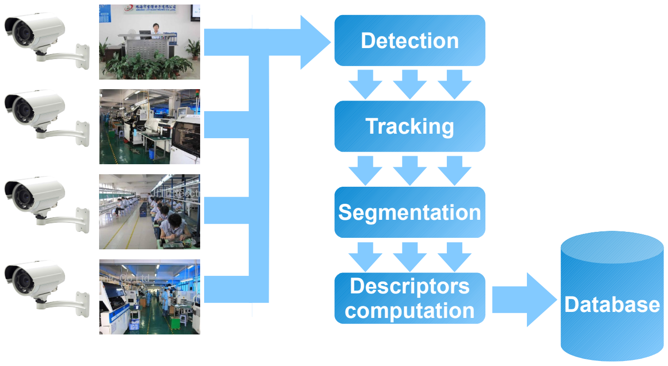


Figure 1.2: Descriptor construction pipeline.

struct descriptors for person re-identification. Clothing appearance exhibits a high degree of uniqueness over limited periods of time, and is relatively easy to extract.

Using the same appearance descriptors above, this thesis also proposes a new useful task, named *appearance-based people search*, which consists of retrieving video sequences of individuals that match a *textual* description of clothing (e.g., "person wearing a black t-shirt and white trousers"). This problem differs from re-identification, as in the latter task the query is an *image* (or a video-sequence), while in the former task the query is a high level textual description. This functionality can be very useful, e.g., in forensics investigations, where a textual description can be provided by a witness. Appearance-based people search bears a close resemblance with *attribute-based people search*, where images of people that show a certain attribute (e.g. the presence of a bag, a certain colour of the shirt) are retrieved. So far, attribute-based search was addressed by very few authors [124, 130].

Whatever is the task of interest (person re-identification or people search), the procedure of extracting appearance descriptors typically follow this pipeline (see Fig. 1.2):

1. the person is *detected* and *tracked* by suitable algorithms;
2. the pixels belonging to the person are separated from the background (*foreground extraction* or *segmentation*) in each frame of the video-sequence;
3. a descriptor is built from the resulting silhouettes (one for each frame), using *local* or *global* features, possibly after different body parts are detected through a body model, in order to take the into account the non-rigid nature of the body;

Descriptors of Step 3 are then stored in a data base for subsequent searches.

Step 1 requires i) a method to detect people in a given video frame [38] (i.e., to recognise the image regions, or *blobs*, that contain a person), and ii) a data association algorithm that track people found by the detector [70, 100, 137] (i.e., to associate blobs in subsequent frames to the same person). These two steps may also be carried out together, and reinforce one another [3]. Step 2 is usually carried out using an adaptive model of the background [44]. Regarding Step 3, a number of different techniques have been proposed in recent years. Often, the body is at first divided in parts; this phase can be carried out using a fixed or adaptive part subdivisions. The whole body, or the parts in which the body has been subdivided, are then described by global features of different kinds (e.g., colour distributions), or by bags (unordered sets) of local features.

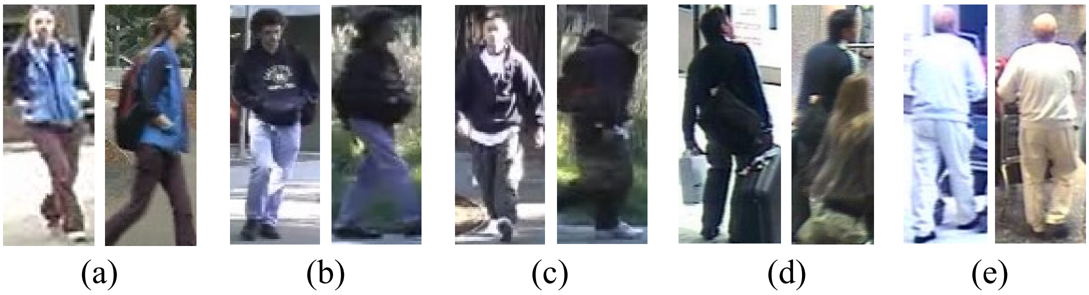


Figure 1.3: Pairs of images showing the same person from different cameras, taken from the VIPeR [57] and i-LIDS data sets [139]. Notice pose variations (a)(b)(c), partial occlusions (d), illumination changes (a)(b)(c), and different colour responses (e).

Many challenging issues can affect some or all the three steps above. Among them we cite (see Fig. 1.3):

- **Pose and viewpoint variations.** The relative pose of a person with respect to the cameras of the network varies depending on the walking path of that person, and of the viewpoint of the camera. This may cause consistent variations of the person appearance.

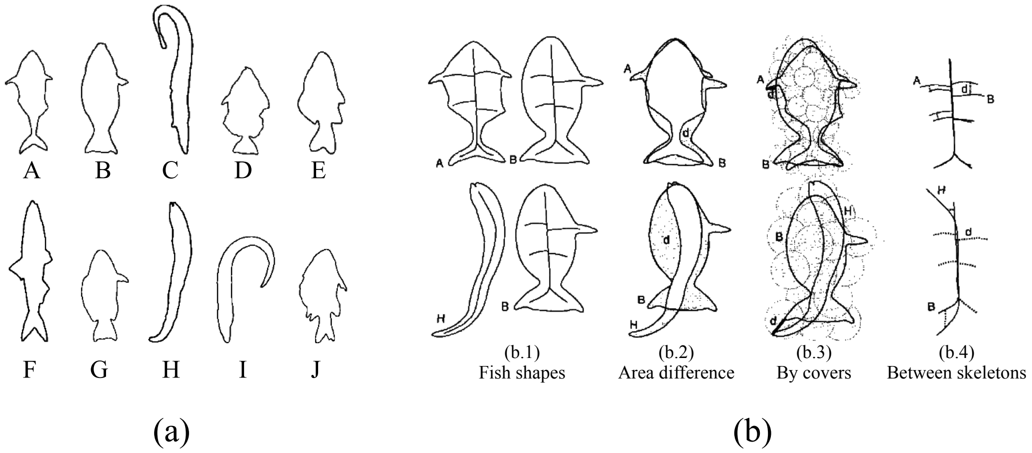
- **Partial occlusions.** Parts of a person may be not visible to the camera due to occlusions caused by objects, clothing accessories or other people. This may cause the segmentation algorithm to fail in separating one person from the rest of the scene; consequently, descriptors may be built from images partially corrupted by the source of the occlusion.

- **Illumination changes.** Illumination conditions may differ in different cameras, and in the same camera in different periods of time due to changing environmental conditions. This may result in appearance changes over different cameras and during time.

- **Different colour responses.** Different cameras may have a different colour response, that may affect person appearance as well.

Various techniques have been put in place to overcome, at least partially, the issues above. Some of them will be discussed in Chapter 2.

Despite various other soft biometrics (e.g., gait) different than the clothing appearance could be in principle used for person re-identification, it is in general difficult or impractical to extract them. Still, the recent introduction of combined video and range (RGB-D) sensors (e.g., MS Kinect [80]), can enable the extraction of some useful soft cues, even in unconstrained environments and without calibration, exploiting the per-pixel depth estimation that is added to the usual RGB information. The Depth information can be used, for instance, to construct a 3D model of the person, or to extract his/her skeleton, that in turn can be used to estimate various anthropometric measures (e.g. the height, the leg length) for re-identification [10]. The combination of anthropometric measures with appearance could in principle lead to a more robust signature of the person and thus to better recognition performance.



(b.1) Fish shapes

(b.2) Area difference

(b.3) By covers

(b.4) Between skeletons

(a)

(b)

Figure 1.4: (a) Fish shapes and (b) various possible approaches to measure their similarity: (b.2) the area of non-overlapping parts; (b.3) one shape is covered by identical balls, such that the balls centres belong to it, taking care that the other shape is covered as well; the shapes are then exchanged and the same procedure is repeated; the sought distance is the radius of the minimal ball; (b.4) Dissimilarity of corresponding skeletons, that can be computed e.g. by summing up the differences between corresponding parts, weighting missing correspondences more heavily. Figures taken from [104].

1.2 Dissimilarity representations

The classical way of representing objects in pattern recognition is to evaluate a set of measurements (*features*) on the object and construct a vector of such measures, commonly called a *feature vector* [42]. Dissimilarity representations have been proposed to deal with classification problems in which a feature vector representation is not available, or it is difficult to find discriminative features, and it is possible to define a dissimilarity measure between pairs of objects, instead [104]. For example, while how to describe the shape of an object may not be clear, it may be simple to define a pair-wise dissimilarity between shapes, e.g. the area difference (see Fig. 1.4 for an example). By means of such measure, it is possible to represent any object with a vector of dissimilarity values to a predefined set of "prototype" objects.

Implicit to the dissimilarity paradigm is the idea that the notion of *proximity* (similarity or dissimilarity) is more fundamental than the notion of *feature* or of *class* [104]. In fact, proximity plays the most important role on the intuitive definition of what constitutes a class, while the way one represents objects (e.g. by feature vectors) is only subsequent. This means that dissimilarity representations are *representation-independent*: in fact, the underlying representation (e.g., feature vectors, if available) of prototypes and objects, is actually not relevant from the point of view of the dissimilarity representation. Fig. 1.5 visualises the differences between feature-based and dissimilarity-based representations.

Prototypes can be chosen in several ways depending on the task at hand, for instance by clustering a given set of objects and taking as prototypes the objects nearest to cluster centroids [104]. The dimensionality of dissimilarity representations is strictly governed by the number of prototypes, as each prototype defines exactly one dimension. How many prototypes are used also usually impacts to the performance of classification algorithms. The number of prototypes becomes therefore an useful parameter for the system designer to gov-

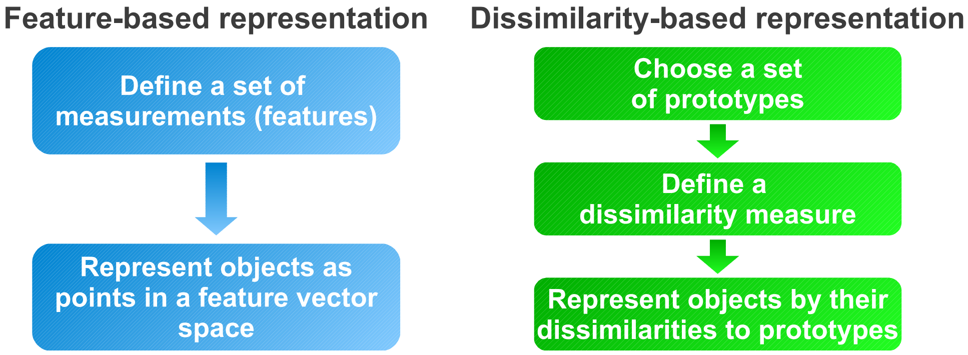

Feature-based representation
Define a set of measurements (features)
Represent objects as points in a feature vector space
Dissimilarity-based representation
Choose a set of prototypes
Define a dissimilarity measure
Represent objects by their dissimilarities to prototypes

Figure 1.5: Feature-based representations and dissimilarity-based representations.

ern the trade-off between the compactness of descriptors (that may reduce computational requirements) and the performance.

An extension of the paradigm for objects models with multiple parts and components

Dissimilarity representations can be exploited in various tasks. In computer vision applications, often objects (e.g., human body) are better described using a part-based model and/or bags of local image features. This is the case, for instance, of typical appearance descriptors used for person re-identification and people search. Therefore, for this kind of applications it could be useful to extend the dissimilarity paradigm so that prototypes can even represent localised image characteristics and refer to specific object parts.

Following this intuition, in this thesis a novel dissimilarity framework is proposed, that extends the original paradigm in two ways: (i) it allows for prototypes to represent either global or local characteristics (both called "components" in the framework's terminology), and (ii) it can associate a specific body part to each prototype. This framework, named Multiple Component Dissimilarity, enables the construction of a dissimilarity representation from any existing appearance descriptor, and even from a combination of descriptors that use heterogeneous body models and features. This novel representation provides a natural way to learn high-level concepts (which enables appearance-based people search). Furthermore, it can drastically reduce the issue of computational complexity of the matching phase of re-identification tasks: in fact, matching two descriptors reduces to comparing small dissimilarity vectors, that is an almost immediate operation with modern CPUs.

Although the framework is inspired by appearance descriptors, its general representation into multiple parts and components is not confined to the task of representing clothings characteristics: indeed, it can embrace other domains. This thesis will preliminarily explore the use of MCD to combine other soft cues with the clothing appearance, when they are available, by combining appearance descriptors with anthropometry for person re-identification.

1.3 Modelling people search on multimedia data

The Multiple Component Dissimilarity framework allows one to cope with two task, person re-identification and people search, in a very similar way. In fact, these tasks are also similar in terms of problem formulation, both being essentially Information Retrieval [91] problems. They differ only for the particular kind of query: an image or a video in the case of person re-identification, and a textual, or semantic, query in the case of people search. Going more general, person re-identification and people search can be seen as specific instances of the problem of retrieving people described by means of any set of modalities (e.g., appearance, anthropometry) extracted from data coming from different media (RGB data, Depth data, but also audio and speech, for instance) with respect to a query which can be expressed at various semantic levels. A query, for instance, of the same kind of the data being processed (e.g. an image), or even at an higher semantic lever (e.g. a textual description). Building on this general idea, in this thesis a possible model of *people search on multimedia data* is proposed, which tries to provide a general framework for such kinds of problems.

1.4 Outline and goals of the thesis

In this introductory Chapter, the reader has been introduced to the motivations and challenges of an important task in intelligent video-surveillance, namely *person re-identification*. The main contributions of this thesis to this fields are:

- a general framework, called Multiple Component Dissimilarity (MCD), that extends the dissimilarity paradigm for pattern recognition to deal with objects consisting of multiple parts and bags of components;
- a method to speed up existing person re-identification methods, that is based on MCD representations and exploits their compactness to reduce computational needs;
- a state-of-the-art re-identification method based on the combination of different kinds of appearance features into one single MCD descriptor;
- a method to combine different descriptors, even heterogeneous and coming from different cues/modalities, into a single, compact one, based on MCD;
- an implementation of the above method to perform person re-identification based on both clothing appearance and anthropometric measures, extracted using RGB-D cameras; to the Author's best knowledge, this is first example of multi-modal person re-identification presented in literature;
- a novel data set for assessing multi-modal person re-identification methods that exploit RGB-D information, made up of RGB and Depth video-sequences showing individuals in different poses and locations, under different illumination conditions;
- a method that uses MCD to perform the new task of "appearance-based people search", by learning high level concepts from dissimilarity representations obtained through MCD;
- a possible formulation of a generalisation of the tasks of people search and person re-identification, that is named "people search on multi-media data".

The rest of the thesis is structured as follows. Chapter 2 provides an overview of existing descriptors for person re-identification. Then, the Multiple Component Dissimilarity framework to construct dissimilarity descriptors and perform person re-identification and people search is presented in Chapter 3. Experimental evidences of the effectiveness of using MCD in the tasks of person re-identification and people search are then provided in the next four Chapters. In Chapter 4, the compactness of MCD-based descriptors is exploited to speed up an existing person re-identification method. In Chapter 5, a novel re-identification method based on MCD is presented, based on the combination of multiple kinds of descriptors into a single dissimilarity vector, and which is able to attain state-of-the-art performance while exhibiting low computational requirements. In Chapter 6, MCD is used to obtain multi-modal person re-identification methods combining clothing appearance with anthropometric measurements extracted through the use of RGB-D cameras. In Chapter 7, is used to implement appearance-based people search. Chapter 8 presents a novel formulation of the problems of people search and person re-identification as instances of a new general task, "people search on multi-media data". Finally, Chapter **??** concludes the thesis, suggesting directions for future research.

Chapter 2

Literature overview

This Chapter provides an overview of existing methods used in literature for the task of person re-identification. As explained in Chapter 1, most methods are based on descriptors of the clothing appearance, which are relatively easy to extract and show a good uniqueness over limited periods of time. Given a frame showing a person, the first step for constructing an appearance descriptor is to extract the pixels that belong to the silhouette of that person. As stated in Sect. 1.1, this is done by i) locating the rectangular region of the image that contains the person, also called *blob* (Fig. 2.1-a), and ii) labelling the pixels of the blob as person or non-person pixels. The second step is to describe the appearance of the person, relying on the pixels belonging to the person's silhouette. This step produces the actual appearance descriptor. Possibly, if more than one frame are available per person (e.g. if the source is a video-sequence), appearance descriptors created from different frames can be conveniently accumulated.

The vast majority of methods assumes that the steps of detection, tracking and segmentation have been already accomplished using any of the algorithms available in literature, and concentrate on the task of constructing descriptors. The interested reader is referred to [38] and [21] for a comprehensive survey of pedestrian detection and foreground segmentation algorithms.

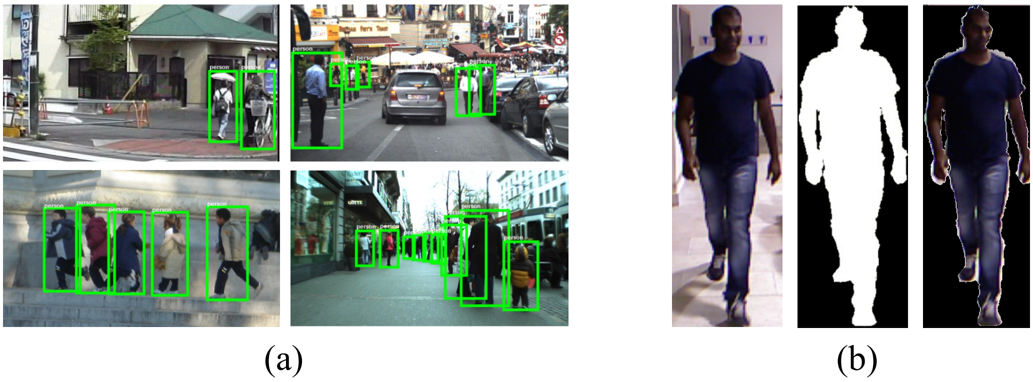


Figure 2.1: (a) Example outputs of a pedestrian detection algorithm in three frames taken from real-world video-surveillance footages; Detected blobs are in green. (b) Example of division of a blob into person and non-person pixels.

Appearance descriptors usually follow a part-based body model: the body is at first subdivided in *parts*, to deal with the non-rigid nature of the human body. Then, body parts are described via global features or bags (i.e., unordered sets) of local features. Body part subdivision models and features used in the literature are described respectively in Sect.2.1 and in Sect.2.2 of this Chapter. Combining different kind of features may help in attaining a better performance; Sect. 2.3 provides a closer insight on typical approaches for feature combination in appearance descriptors.

While almost all existing methods use the clothing appearance as main cue to perform re-identification, it is worth to note that other approaches have been attempted in literature, for instance based on gait, or anthropometric measures captured through novel RGB-D sensors. These methods are briefly surveyed in Sect. 2.4, which concludes this Chapter.

2.1 Part-based body models

The human body is not a rigid object. Instead, it has a complex kinematics, and can be better described by a part-based model, possibly where relative positions of parts are not fixed a-priori but are inferred from the image. Furthermore, discontinuities of the clothing appearance usually follow the body structure (e.g., the clothing appearances of the upper and lower body usually differ). Many existing appearance descriptors, therefore, exploit some part-based human body model to segment the silhouette into different parts. Some other descriptors (e.g., [7, 13, 20, 33, 58, 63, 64, 68, 73, 77, 88, 90, 107, 126, 127]) consider the body as a whole instead. Part-based body models used in existing appearance descriptors can roughly be divided into three categories:

- *fixed models*, in which size and position of body parts are chosen a-priori by the designer;
- *adaptive models*, that try to fit a predefined part subdivision model to the image of the individual;
- *learned models*, that use a part-based body model that is previously learnt from a training set of images of individual.

In the rest of the Section, part-based body models belonging to the three categories above are reviewed and compared.

2.1.1 Fixed part models

Probably the simplest kind of part subdivision is a fixed one, in which the sizes and positions of body parts are chosen a-priori. An example of this approach can be found in [86, 110, 140], where the body is subdivided into six horizontal stripes of equal size, that roughly capture the head, upper and lower torso and upper and lower legs. Similarly, in [6] the silhouette is subdivided in five equal-sized stripes. An even simpler fixed part subdivision is used in [78]. Three horizontal stripes of respectively 16%, 29% and 55% of the total blob height roughly locate head, torso and legs, then the first strip is discarded as the head typically consists of few pixels and is not informative for the clothing appearance.

2.1.2 Adaptive part models

Other body models are *adaptive*, in the sense that they try to fit a predefined part subdivision model to the image of the individual. In one of the descriptors proposed in [8], the MPEG-7 Dominant Colour Descriptor (DCD) [136] is used to dynamically separate the body into two parts, upper and lower body, looking for discontinuities in dominant colours (the same DCD is also used as feature set to describe each body part, see Sect. 2.2). The approach of [45] extends the basic idea of exploiting appearance anti-symmetries of [8]. It dynamically finds three body areas, namely the head, torso, and legs, exploiting symmetry and anti-symmetry properties of silhouette and appearance. To this aim, two operators are defined. The first measures is called *chromatic bilateral operator*. It measures the appearance anti-symmetry of a certain image region with respect to a given horizontal axis, and is defined as

$$C(y,\delta) = \sum_{B_{[y-\delta,y+\delta]}} d^2\left(p_i, \hat{p}_i\right), \tag{2.1}$$

where $d(\cdot,\cdot)$ is the Euclidean distance, evaluated between pixels represented in the HSV colour space p_i and $\hat{p}_i$ located symmetrically with respect to an horizontal axis placed at height y of the person image. This distance is summed up over the person pixels lying in the horizontal strip $B_{[y-\delta,y+\delta]}$ centred in y and of height 2δ.

The second is called *spatial covering operator* and measures the difference of the silhouette areas of two regions:

$$S(y,\delta) = \frac{1}{W\delta}\left|A\left(B_{[y-\delta,y]}\right) - A\left(B_{[y,y+\delta]}\right)\right|, \tag{2.2}$$

where W is the width of the blob, and $A\left(B_{[y-\delta,y]}\right)$ and $A\left(B_{[y,y+\delta]}\right)$, denote the number of person pixels respectively of the strip of vertical extension $[y-\delta, y]$ and $[y, y+\delta]$. These operators are combined to find two axes, y_{HT} and y_{TL}, that respectively separate head and torso, and torso and legs. These axes are defined as

$$y_{TL} = \arg\min_{y}\left(1 - C(y,\delta) + S(y,\delta)\right), \tag{2.3}$$

$$y_{HT} = \arg\min_{y}\left(-S(y,\delta)\right). \tag{2.4}$$

The parameter δ is set to a value of $\delta = Y/4$ where Y is the blob height in pixels. The values y_{HT} and y_{TL} isolate three regions approximately corresponding to head, body and legs (Fig. 2.2-a). The head part is discarded as it carries very low informative content. As claimed by the authors, this strategy is able to locate body parts which are dependent on the visual and positional information of the clothes, robust to pose, viewpoint variations, and low resolution. After [45], the same part-based model has been used in various other works [14, 92, 93, 133].

A deformable model that is fitted to each individual to find six body regions is used one of the methods in [55], based on decomposable triangulated graphs [2]. A triangulated graph is a collection of cliques of size three, that has a perfect elimination order for their vertices, i.e., there exists an elimination order for all vertices such that (i) each eliminated vertex belongs only to one triangle, and (ii) a new decomposable triangulated graph results from eliminating the vertex.

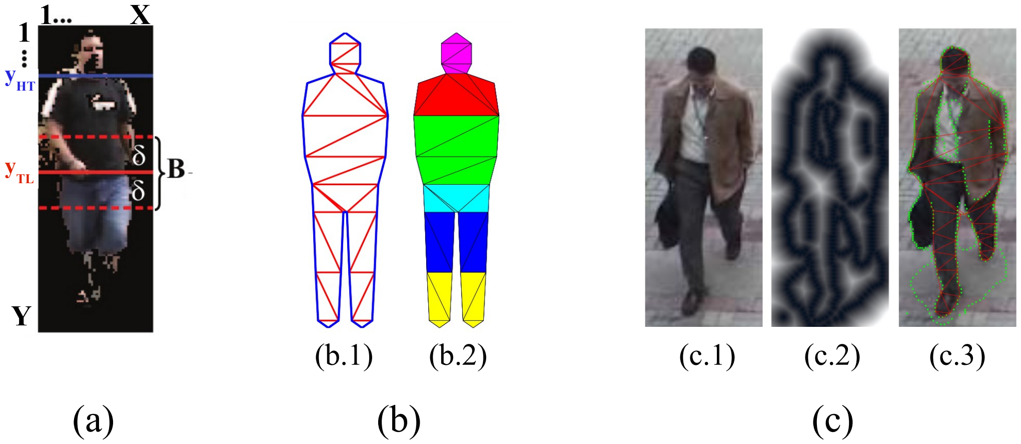


Figure 2.2: (a) Symmetry-driven subdivision in three parts [45]. The blob of size $Y \times X$ pixels containing the person is divided according to two horizontal axes, y_{HT} and y_{TL}, found by minimising a proper combination of the operators defined in Eqs. (2.1)-(2.2). (b) Decomposable body model used in [55]: (b.1) the decomposable triangulated graph model; (b.2) Partitioning of the person according to the decomposable model. (c) An example of fitting the decomposable triangulated model of [55] to an individual: (c.1) an image of an individual; (c.2) edges detected through the the Canny's algorithm [25]; (c.3) result of fitting the model to the edges (in red). All figures are taken from [45] and [55].

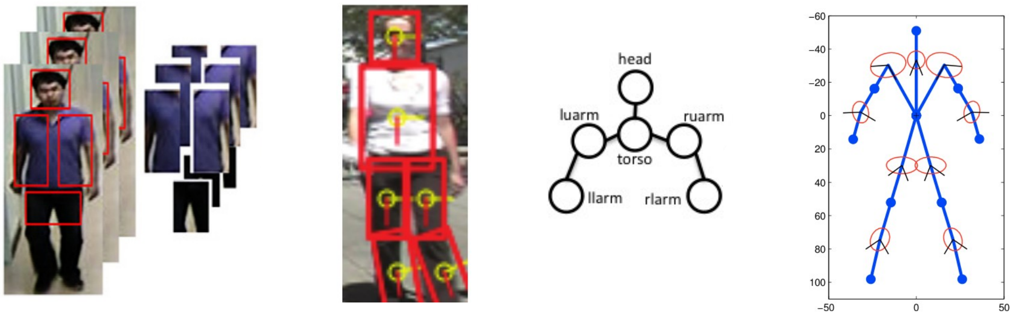
The model is fit to the image of a person using the following strategy. Let the model be a decomposable triangulated graph T with n triangles $T_i, i = 1, \ldots, n$. The goal is to find a function g that maps the model to the image domain, such that the consistency of the model with salient image features is maximised, and deformations of the underlying model are minimised. The function g must be a piecewise affine map [47], i.e the deformation of each triangle $g_i(T_i)$ must be an affine transformation. The problem becomes to minimise an energy functional $E(g, I)$ that can be written as a sum of costs:

$$E(g, I) = \sum_i E_i(g_i, I) = \sum_i \left(E_i^{data}(g_i, I) + E_i^{shape}(g_i) \right), \tag{2.5}$$

where the I represents the image features. The terms $E_i^{shape}(g_i)$ take into account the cost for shape distortion of the i-th triangle, while $E_i^{data}(g_i, I)$ attracts the model to salient image features, which are found using an edge detector (Canny's algorithm [25]). As shown in [2], a model based on decomposable triangulated graphs can be efficiently optimised using dynamic programming. Once the model has been fitted with regard to the image, the individual is partitioned into six salient body parts, shown Fig. 2.2-b with different colours. An example of application to a real pedestrian image is shown in Fig. 2.2-c.

2.1.3 Learned part models

More recently, some methods that rely on previously trained body part detectors and articulated body models have been proposed. Part detectors are statistical classifiers that learn a model of a certain body part (e.g., an arm) from a given training set of images of people where body parts are manually located and labelled. Typically, these detectors exploit features related to the edges contained on the image. An approach of this kind has been used in



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

Figure 2.3: (a) Sample output of the articulated body model used in [15, 16]. (b) Sample output of the Pictorial Structure model used in [28]. (c) Sample Pictorial Structure of the upper body part, with the torso part as root node. (d) Kinematic prior learned on the dataset from [111]. The mean part position is shown in blue dots; the covariance of the part relations in the transformed space is shown using red ellipses. Figures taken from [15] and [4].

[16, 15] based on the work of Felzenszwalb et al. [46]. The overall body model is made up of different part models; each one, in turn, consists of a *spatial model* and of a *part filter*. The spatial model defines a set of allowed placements for a part with respect to the bounding box containing the person, and a deformation cost for each placement. To learn a model, a generalisation of Support Vector Machines (SVM) [23] called latent variable SVM (LSVM) is used. In [16, 15], such model is used to detect four different body parts, namely head, left torso, right torso and the upper legs (see Fig. 2.3-a).

An articulated body model based on Pictorial Structures (PS) was proposed in [4] and later exploited in [28] for the task of re-identification. In [28], six parts are considered (chest, head, thighs and legs, see Fig. 2.3-b), while the original PS model is also able to detect and locate upper and lower arms.

A PS model for an object [48] is a collection of parts with connections between certain pairs of parts (an example is provided in Fig. 2.3-c). The approach of [4] uses a PS of the human body that is made up of a set of N parts, and a set of generic part detectors based on descriptors of the shape. The model and the body part detectors are trained on a training set of images of people.

Let $L = \{\mathbf{l}_0, \ldots, \mathbf{l}_{N-1}\}$ be the set of configurations of each body part. Each $\mathbf{l}_i$ is the *state* of the i-th body part $\mathbf{l}_i = (x_i, y_i, \theta_i, s_i)$, where x_i and y_i are the image coordinates of the part centre, θ_i is the absolute part orientation, and s_i is the part scale, relative to the size of the part in the training set. Given the image evidence D, the problem is to maximise the a-posteriori probability (*posterior*) $p(L|D)$ that the part configuration L is correct. The posterior is proportional to

$$p(L|D) \propto p(D|L)p(L) \tag{2.6}$$

according to Bayes' theorem [42]. The term $p(D|L)$ is the likelihood of the image evidence given a particular body part configuration, while $p(L)$ corresponds to a kinematic tree prior. Both are learned from a training set, as follows.

Kinematic three prior. The prior $p(L)$ encodes the kinematic constraints, i.e. the constraints on the relative parts disposition. The body structure is mapped on a directed acyclic graph, so that $p(L)$ can be factorised as

$$p(L) = p(\mathbf{l}_0) \prod_{(i,j)\in E} p\left(\mathbf{l}_i|\mathbf{l}_j\right) \tag{2.7}$$

where E denotes the set of all directed edges in the kinematic tree, and $\mathbf{l}_0$ is the root node, that in [4] is chosen to be the torso body part.

The prior for the root part configuration $p(\mathbf{l}_0)$ is assumed to be uniform. To model part relations $p(\mathbf{l}_i|\mathbf{l}_j)$, a transformed space is used, where such relations can be modelled as Gaussian [48]. More specifically, the part configuration $\mathbf{l}_i = \left(x_i, y_i, \theta_i, s_i\right)$ is transformed into the coordinate system of the joint between the two parts i and j using the transformation:

$$T_{ji}(\mathbf{l}_i) = \begin{pmatrix} x_i + s_i d_x^{ji} cos\theta_i + s_i d_y^{ji} sin\theta_i \\ y_i + s_i d_x^{ji} sin\theta_i + s_i d_y^{ji} cos\theta_i \\ \theta_i + \bar{\theta}_{ji} \\ s_i \end{pmatrix} \tag{2.8}$$

where $d^{ji} = \left(d_x^{ji}, d_y^{ji}\right)^T$ is the mean relative position of the joint between the two parts i and j, in the coordinate system of part i, and $\bar{\theta}_{ji}$ is the relative angle between the two parts. Then, part relations are modelled as Gaussian in the transformed space:

$$p\left(\mathbf{l}_i|\mathbf{l}_j\right) = \mathcal{N}\left(T_{ji}(\mathbf{l}_i)|T_{ij}(\mathbf{l}_j), \Sigma^{ji}\right) \tag{2.9}$$

where d^{ji} and Σ^{ji} can be learned via maximum likelihood estimation [42] from a labelled training set of images of people. It is worth noting that the body parts are only loosely attached to the joints (also called a *loose-limbed model* [120]), which helps increasing the robustness of the pose estimation. Fig. 2.3-d shows the priors learned from the multiple views and multiple poses people data set of [111], a common benchmark corpus for body pose estimation algorithms.

Likelihood of the image evidence. To estimate the likelihood $p(D|L)$, the methods relies on a different appearance model for each body. Each appearance model will result in a part evidence map $\mathbf{d}_i$ that reports the evidence for the i-th part for each possible position, scale, and rotation.

Assuming that the different part evidence maps are conditionally independent, and that each $\mathbf{d}_i$ depends only on the part configuration $\mathbf{l}_i$, the likelihood $p(D|L)$ can be written as:

$$p(D|L) = \prod_{i=0}^{N} p\left(\mathbf{d}_i|\mathbf{l}_i\right). \tag{2.10}$$

Substituting Eq. (2.7) and Eq. (2.10) in Eq. (2.6), one finally obtains:

$$p(L|D) \propto p(\mathbf{l}_0) \cdot \prod_{i=0}^{N} p\left(\mathbf{d}_i|\mathbf{l}_i\right) \cdot \prod_{(i,j)\in E} p\left(\mathbf{l}_i|\mathbf{l}_j\right) \tag{2.11}$$

The part detectors $p\left(\mathbf{d}_i|\mathbf{l}_i\right)$ use a variant of the shape context descriptor [95], that consists in a log-polar histogram of locally normalised gradient orientations. The feature vector is obtained by concatenating all shape context descriptors whose centres fall inside the bounding box of the part. During detection, different positions, scales, and orientations are scanned with sliding windows. The classifier used for detection is an ensemble of a fixed number of decision stumps combined through AdaBoost [52].

2.2 Features

Each body part (or the whole image of the individual, if no body part subdivision model is used) is typically described using one or more different global or local features. In this, Section, the main kinds of features used in the literature are reviewed.

2.2.1 Global features

Global features are characteristics measured in the whole image or body region considered, and are usually represented as a fixed-size vector of real numbers.

Probably the most widely used feature of this kind is the global colour histogram. Given a colour image of size $N = W \times H$ pixels, the colours of the image are at first quantised into B bins $1,\dots,B$. The histogram is then constructed as the count of the number of occurrences per bin. Typically, such count is normalised as the fraction of pixels of the image belonging to the bin. Colour image pixels are typically represented as a triplet of values, representing the amount of colour in different colour channels (e.g., Red, Green and Blue). In this case, each colour channel is quantised separately. The resulting histogram can be multi-dimensional (one dimension for each channel), or mono-dimensional (the final histogram is constructed as the concatenation of histograms in each colour channel). The latter saves a lot of space (e.g., if 16 bin are used for each colour channel, the size of the multi-dimensional histogram would be $16 * 16 * 16 = 4096$ bins, while the mono-dimensional one would have a size of 48 bins) and has usually a similar discriminant capability to the former. Various colour spaces exist in the literature. Among them it is worth citing:

- The RGB colour space, where each colour is represented as the corresponding amount of Red, Green and Blue; it directly relates to the way devices acquire and visualise colours.

- *Perceptual* colour spaces, i.e., spaces inspired to the way the human brain perceives colour; e.g., the Hue-Saturation-Value (HSV) colour space, in which the light intensity (V channel) is separated from the colour tonality (H channel) and the saturation of the colour (S channel).

Good surveys on colour spaces are provided in [125, 129]. Many appearance descriptors use global colour histograms, to represent the whole body appearance [13, 73, 86] or the overall appearance of each body part [6, 14, 15, 16, 45, 55, 58, 78, 110, 133, 140]. Du et al. [40] recently evaluated histograms computed in various colour spaces for building appearance descriptors for person re-identification. To tackle with the lower amount of information usually carried by peripheral pixels (that could actually belong to the background, as the person segmentation is usually very noisy), in [28, 45, 133] these pixels receive less weight than those near the vertical silhouette symmetry axis.

The colour space is typically quantised in an uniform fashion. However, many colour ranges can be irrelevant for representing a certain appearance, e.g. colours ranges that are not present in the image, or whose coverage percentage with respect to the image is irrelevant. For this reason, some approaches try first to find the most representative colour ranges, then describe the appearance with respect to these ones. One of the methods of [8] and the methods of [15, 16, 77] use the Dominant Colour Descriptor (DCD) (also called Representative Meta Colours Model, RMCM) of MPEG-7, which provides a compact description of the most representative colours. Given an image, the DCD algorithm first finds the K *dominant colours* [36], via *k-means* clustering of all the colour triplets in the image. Then, the descriptor is defined as

$$F = \{\{c_i, p_i\}, i = 1, \ldots, K\} \tag{2.12}$$

where c_i is the i-th dominant colour (i.e., the centroid of the i-th cluster), and p_i is the percentage of image pixels that fall into the i-th cluster. A similar approach is used also in [24], called Global Colour Context. The method of [33] partly differs to the former ones, although it shares with them the same idea of describing appearance in terms of the most important colours. Instead of finding representative colours by clustering, they are chosen a priori; specifically, eleven colors, usually referred to as *culture colours* [32], are used: black, white, red, yellow, green, blue, brown, purple, pink, orange, and grey. Each pixel of the image is assigned to the most similar cultural colour.

Colour histograms are invariant to scale and show a good robustness with respect to partial occlusions, if the occlusion itself is small. However, they are sensitive to changing brightness and colour response of the sensor. Illumination conditions in outdoor environments may consistently vary during time due to changing weather conditions and the varying illumination of the Sun during the day. On the other hand, lighting conditions of indoor scenes may vary from camera to camera due to different types of lamps (e.g., incandescent, tungsten, neon) and also due to weather conditions in case of presence of windows that let the Sun light enter. Colour response of the sensors may also vary due to environmental conditions and due to the automatic colour balance that often takes place in-camera.

Different mechanisms have been exploited to tackle with the above problems. Probably the simplest one is colour normalisation [129]. The chromaticity RGB space is one of these techniques, used in [20, 40, 127], and consists of dividing each colour channel of each pixel by the sum of all the channels of that pixel, e.g. $R' = R/(R+G+B)$. Another common technique is the *Grey-world normalisation* [22], which relies on the assumption that the average colour of a scene is usually a tonality of grey. It consists of dividing each RGB channel of every pixel by the average value of that channel in the image, e.g. $R' = R/\mathrm{mean}(R)$. Grey-world normalisation is used in [126, 127]. Similar to Grey-world is the affine normalisation used in [20, 126, 127], where pixel-values of each color channel are normalised independently by subtracting the average and scaling them with the standard deviation, e.g. $R' = \big(R - \mathrm{mean}(R)\big)/\mathrm{std}(R)$.

Alternative to colour normalisation is histogram equalisation [49], which is used in the re-identification methods of [9, 126, 127]. It is based on the assumption that a change in illumination preserves the rank ordering of sensor responses (i.e. pixel values). The rank measure for the $i-th$ bin of the histogram and the k-th colour channel is defined as $M_k(i) = \sum_{u=0}^{i} H_k(u) / \sum_{u=0}^{N} H_k(u)$, where N is the number of bins and $H_k()$ is the histogram relative to the k-th channel.

Finally, Piccardi and Cheng [107] exploited a colour quantisation scheme to mitigate the effect of illumination changes between cameras. They represent the image with a Major Colour Spectrum Histogram (MCSH), that is, an histogram of the top N represented colour values in the image.

Another problem of histograms is that they do not retain any information on the spatial disposition of colours. A simple way to incorporate the spatial information is to add the relative pixel height (i.e. the ratio between the vertical coordinate of the pixel and the total height of the silhouette) as another channel of the image[1]. A colour-position histogram can be then built which is able to spatially localise the colour distribution [20, 127, 126]. A similar approach is used also in [78], where two dimensions are added to each pixel (i.e. the radial and angular distance to the torso center) and quantised. The Color Structure Descriptor (CSD) of MPEG-7 [90] is used in [62], and encodes the distribution of colour by the following steps: (i) move a window of size 8×8 pixel over the picture ; (ii) determine which colours are present in within the window; (iii) increase the corresponding bins in a color histogram by one, independently of the number of pixels of these colors.

Instead of looking at colour properties, other kinds of global features try to characterise gradients, textures and repeated patterns of the whole body appearance or of each body part. Gabor filters [97] ans Schmid filters [117] are orientation-sensitive filters that capture texture and edge informations on the image. The former ones are aimed at detecting horizontal and vertical lines, while the latter ones detect circular gradient changes. They are used in various appearance descriptors [58, 86, 88, 110, 140] in conjunction with other colour-related features.

Hahnel et al. [62] compared various different texture features. The fist is the 2D Quadrature Mirror Filter (QMF), a well known filter in signal processing that splits a 2D input signal into two bands (high and low-pass) in each direction (horizontal, vertical and diagonal. The second is the Oriented Gaussian Derivatives (OGD) filter, based on steerable Gaussian filters. Also, two MPEG-7 texture-related descriptors, are used the Homogeneous Texture Descriptor (HTD) that uses Gabor filters, and the Edge Histogram Descriptor (EHD), basically an histograms of the directions of each edge pixel in the image [121].

It is worth pointing out that texture-based features have always been used in combination to colour-based ones. Information on repeated patterns is in fact likely to be not distinctive enough when used alone. Hahnel et al. [62] confirmed this thought, and showed also that the combination of colour and texture-based descriptors may lead only to minor performance improvements.

2.2.2 Local features

The term *local feature* refers to an appearance characteristic of a small portion of the image (e.g., the neighbourhood of a pixel). The regions where local features are extracted can be chosen in various way (e.g. by dense sampling, by an interest operator or at random). Each small region is described by a feature vector (e.g., an histogram). This lead to a representation of the image as as a *bag* (set) of local features.

Interest points are one important category of local features. The most famous among them is SIFT (Scale Invariant Feature Transform) [87], where at first salient points of the image are chosen via in interest operator that looks for “stable” locations in the image (i.e. locations that are identifiable over different scales and rotations). This operation is carried out by detecting scale-extrema locations in the *scale space* of scale σ, which is defined by the function

$$L(x, y, \sigma) = \mathcal{N}(x, y, \sigma) * I(x, y) \tag{2.13}$$

where $*$ is the convolution operation in the image coordinates x and y, and $\mathcal{N}(x, y, \sigma)$ is a 2-D Gaussian with standard deviation σ. Stable key-points can be detected in this space e.g.

[1] The horizontal coordinate of the pixel is typically not used, as it is not robust to body rotations and viewpoint changes.

by using difference-of-Gaussians functions convolved with the image:

$$D(x,y,\sigma) = \big(\mathcal{N}(x,y,k\sigma) - \mathcal{N}(x,y,k\sigma)\big) * I(x,y) = L(x,y,k\sigma) - L(x,y,\sigma) \tag{2.14}$$

To detect the local minima and maxima of $D(x,y,\sigma)$, each point (x,y) is compared with its 8 neighbours at the same scale $k\sigma$, and its 9 neighbours in the two scales $(k-1)\sigma$ and $(k+1\sigma)$. If this value is the minimum or maximum of all these points, then this point is an extrema, and it is labelled as key-point. A subsequent stage filters out low-contrast and noisy points. The remaining key-points are described as a histogram of the edge orientations of a small window centred on the key-point. SIFT points or its variants, (e.g., Speeded-Up Robust Features, SURF [12]) are used in various appearance descriptors [35, 63, 64, 77, 88, 92, 93] to represent the whole body appearance.

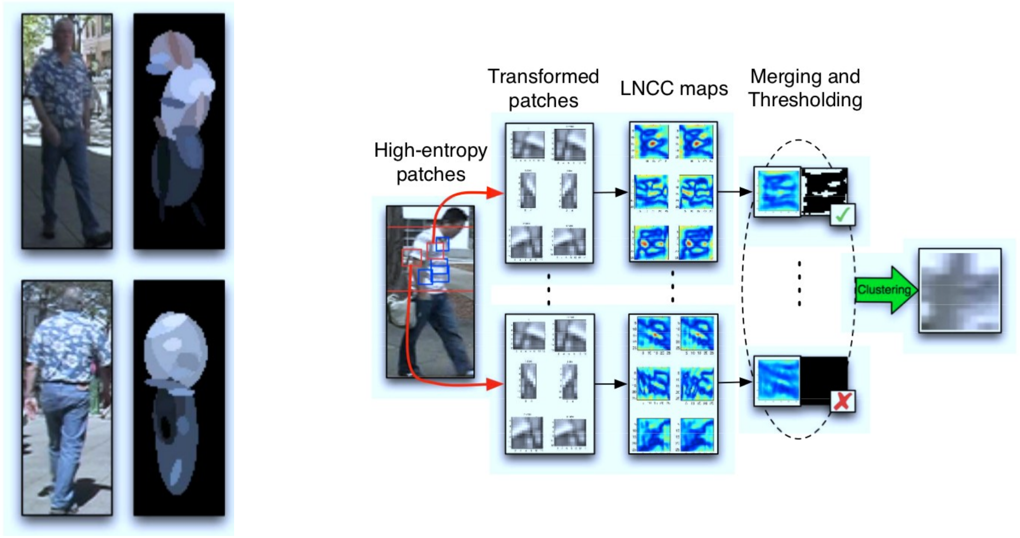
Other approaches use different kinds of local features. Maximally Stable Colour Regions (MSCR) [51] are used in [28, 45, 88]. The MSCR algorithm first detects a set of regions in the image (Fig. 2.4-a) by using a constrained agglomerative clustering on image pixels, which show the maximal chromatic distance. The detected regions are then described by their area, centroid, second moment matrix and average color, forming 9-dimensional feature vectors, and are stable to scale and affine transforms.

Recurrent Highly-Structured Patches (RHSP) used in the method of [45], try instead to capture repeated patterns and textures of the clothing appearance. The procedure of creating RHSPs is as follows. First, random and possibly overlapping small patches are extracted from the image. Patches that do not carry texture informations (e.g. showing uniform colours) are discarded by thresholding the patch entropy, computed as the sum of the entropy of each colour channel. Remaining patches are then further filtered, keeping only those that exhibit invariance to rotations. Second, the recurrence of each patch is evaluated, via Local Normalised Cross-Correlation over a small local region containing that patch. Third, patches that show a high degree of recurrence are clustered, maintaining for each final cluster the patch nearest to the centroid. These patches are finally described as their Local Binary Pattern histogram [102], a simple yet efficient way to describe textured content, based on a per-pixel transform that encodes small-scale appearance structures.

Instead of using interest operators or proper selection criteria to choose where to extract a local feature, in [58], a set of strips of fixed height and position are extracted from the image, and described by a concatenation of colour histograms in different colour spaces and Gabor and Shmid filters. Similarly, in [68] partly overlapping rectangular patches of fixed size are sampled from the image following a pre-defined regular grid. Each patch is represented by its colour histogram in the HSV colour space, and by its LBP histogram to capture textures and repeated patterns.

2.3 Combination of features and matching

Many person re-identification methods use appearance descriptors made up of only one kind of features among the above mentioned ones, typically based on colour or interest points [6, 8, 15, 16, 20, 24, 33, 55, 63, 64, 78, 92, 93, 126, 127]. However, as combining different sources of information usually helps in attaining a better performance, especially when sources are complementary (i.e. they look at different aspects of the appearance, e.g. colour and texture), many authors have defined descriptors that use a combination of features.



(a) (b)

Figure 2.4: (a) Maximally Stable Colour Regions [51] detected in two images showing the same pedestrian. (b) Steps of the extraction of RHSP: random extraction, rotational invariance check, recurrence check, entropy thresholding, clustering. The final result of this process is a set of patches (in this case only one) characterising repeated patterns of each body part of the individual. Figures taken from [45].

In principle, two main combination techniques can be exploited to this aim [114]:[2]

1. *feature*-level fusion: if the features used are made up of a single vector of fixed size (e.g. global features, or local features with an intrinsic ordering) they can be combined simply by concatenating feature vectors;

2. *score*-level fusion: a distinct detector/matcher is used for each feature, and their real-valued scores are combined (e.g., by averaging them, or using their maximum value).

The first approach is followed for instance in [40, 68, 133]. The second approach requires to define a proper fusion rule. Many methods used a weighted average of the partial scores attained with each single feature, where weights are fixed a-priori by the system designer [13, 14, 28, 45]. Another approach is to learn a proper metric or a set of weights from a training set. In [58], AdaBoost[52] is used to this aim: each feature set is associated to a weak two-class classifier (a decision stump) which discerns between the class 0 (identities differ) and 1 (identity is the same) based only in that feature set. The method of [110] tries to find a linear function to weight the absolute difference of samples by training an ensemble of RankSVM rankers [75] given pairwise relevance constraints. The Probabilistic Relative Distance Comparison (PRDC) technique of [140] maximises the probability that a pair of true match has a smaller distance than that of a wrong match. The output is an orthogonal matrix which essentially encodes the global importance of each feature. In [88] a pairwise metric is learned through a recently proposed method, Pairwise Constrained Component Analy-

[2] In *verification* tasks, whose goal is to establish whether the claimed identity is true, combination can also be performed at *decision* level, i.e., by combining the crisp outputs of classifier/detectors. It can not be applied to person re-identification, which is a *recognition* task instead.

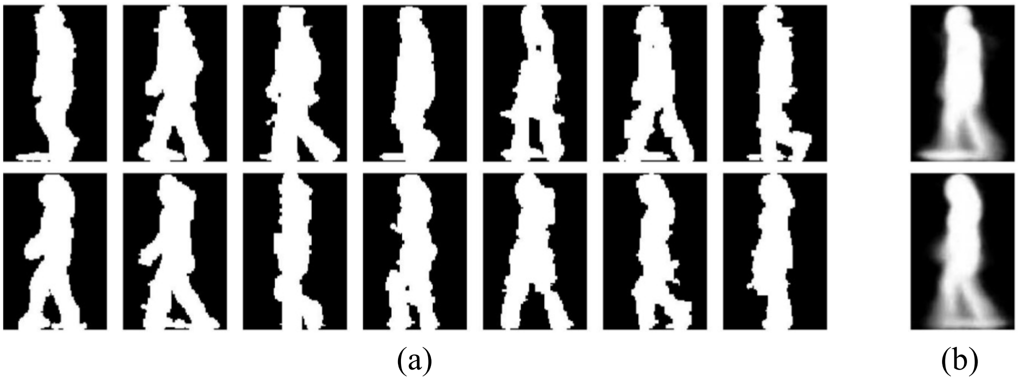


Figure 2.5: (a) Two sequences of aligned foreground silhouettes. (b) Their corresponding Gait Energy Image. Figures taken from [65].

sis (PCCA) [94], which learns a projection into a low-dimensional space where the distance between pairs of data points respects the desired constraints.

Metric learning and similar approaches always help in boosting re-identification performance. However, it is worth to note that all the above methods require a training set of labelled data, which is usually. Such set can be for instance the gallery of templates. This requires that the template gallery is *fixed*, i.e. templates cannot be added during system operation. Obviously, this constraint is usually too strong for many real-world application scenarios.

2.4 Other cues

Some cues alternative to the clothing appearance have been exploited in the literature to perform person re-identification or assimilable tasks. Despite the intrinsic limitations of such cues, they could be potentially of help in certain conditions, possibly combined with appearance cues.

Human *gait*, i.e. the recurrent pattern of motion of a person walking, is among these cues. In cognitive science, it is known to be one of the cues that humans exploit to recognise people [122]. Among the approaches to characterise gait, the recently proposed Gait Energy Image (GEI) [65] has attracted the attention of many researchers. Here, the gait signature is formed by by normalising, aligning and averaging a sequence of foreground silhouettes corresponding to one “walking period” (see Fig. 2.5). Principal Component Analysis (PCA) is then used to reduce the dimensionality of the signature.

The use of Gait Energy Image can lead to high recognition rates [134] and can overcome one of the main limitations of clothing appearance-based approaches, that is, the impossibility of distinguishing people when their clothing changes between observations. It is also not directly affected by illumination changes. However, it requires perfect alignments of the silhouettes to be compared, and is sensible to segmentation errors. These two constraints severely limit the use of GEI-based methods on practical, real-world applications. Researchers have therefore attempted to explore other approaches. Zhao et al. [138] and more recently Gu et al. [59] used a 3D skeletal representation, that however requires multiple overlapping camera views or a constrained environment to construct and track it.

Some authors attempted instead to perform *remote face recognition* [99], that is, face recognition with low resolution images. As low resolution face images are not directly usable for recognition, many approaches attempted to address the problem through the obvious way of trying to increase image resolution, using super-resolution techniques [60, 66, 74, 118]. Some authors proposed instead techniques that work directly on low resolution images, by exploiting metric learning [85, 84], multidimensional scaling [19], or multiple frames from video sequences [5]. All the approaches above could in principle be used in conjunction with appearance cues to increase re-identification accuracy when the face is visible.

Another useful set of soft cues is anthropometry, that is, the characterisation of individuals through the measurement of physical body features [113], e.g., height, arm length, and eye-to-eye distance. Measures are typically taken according to a number of body landmark points (e.g., elbows, hands, knees, feet), that have to be localized either automatically or manually. In the classic study by Daniels and Churchill [34], the uniqueness of 10 different anthropometric traits was evaluated on a large data base of 4063 individuals. None of the considered traits was found to be "average" (i.e., approximately close to the mean point), considering all 10 dimensions. Furthermore, only 7% of the individuals were "average" in 2 dimensions, and 3% in 3 dimensions.

Although the use of anthropometric measurements for person recognition has been proposed in many works, their extraction was often based on costly devices, like 3D laser scanners, and/or require user collaboration in a constrained environment [101, 56, 98]. In some works, anthropometric measurements are extracted from a single RGB camera view, instead. In [11] a method that does not require camera calibration was proposed, for simultaneously estimating anthropometric measurements and pose. However, the former are measured up to a scale factor, and consequently can not be used to directly compare individuals in images acquired by different cameras. Calibration is not required in [1] as well, although 13 body landmarks have to be manually selected, from an image of an individual in frontal pose. Other methods focus on height measurement only [89, 82, 17, 53, 83], but require camera calibration to estimate absolute height values. Interestingly, in [89] height is used as a cue for the task of associating tracks of individuals coming from disjoint camera views, which is actually the same *re-acquisition* task that is enabled by person re-identification.

None of the above works fits the typical setting of person re-identification tasks, which is characterised by multiple, uncalibrated cameras and unconstrained environment, with free poses and non collaborative users. Recently, it has been shown that body pose can be reliably estimated in real-time by exploiting RGB-D sensors [119, 123], like the MS Kinect, a device recently introduced in the video-gaming market. The pose estimation functionality of Kinect SDK [80], which is based on a similar method, provides the absolute position (in meters) of 20 different body joints in real-time, with high reliability (see Fig. 2.6). Detecting joint positions enables the evaluation of several anthropometric measures. In [10] such joints were used to extract a set of different anthropometric measures from front or back poses: distance between floor and head, ratio between torso and legs, height, distance between floor and neck, distance between neck and left shoulder, distance between neck and right shoulder, and distance between torso center and right shoulder. Other three geodesic distance measures were estimated from the 3D mesh of the abdomen, obtained from the Kinect depth map: torso center to left shoulder, torso center (located in the abdomen) to left hip, and between torso center to right hip. Results reported in [10] appear promising. However, many of the considered anthropometric measures are hard or impossible to extract from unconstrained poses. For instance, extracting measures from 3D mesh requires near-frontal pose (abdomen is hidden in back pose); neck distance to left and right shoulders

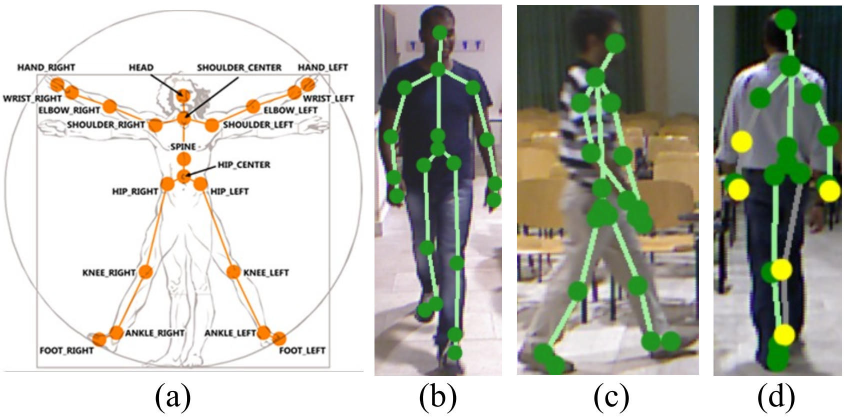


Figure 2.6: (a) The 20 skeletal points tracked by the Kinect SDK in the classical representation of the Vitruvian Man. (b–d) Examples of the pose estimation capabilities of the Kinect SDK. Depending on the degree of confidence of the estimation of the points position, the Kinect SDK distinguishes between *good* (in green) or *inferred* (in yellow) points, the latter being less reliable than the former.

becomes hard to compute from lateral pose, even using a depth map, and requires to distinguish between left and right body parts. Such issues limit the actual set of anthropometric measures that can be used in realistic scenarios.

Chapter 3

The Multiple Component Dissimilarity framework

This Chapter describes the Multiple Component Dissimilarity (MCD) framework and its application to person re-identification and people search. The framework extends the dissimilarity paradigm for pattern recognition, originally proposed by Pekalska and Duin [104]. In particular, MCD aims at representing in a dissimilarity space objects that are made up of multiple *parts*, and that are better described taking into account localised characteristics. This class of objects includes, for instance, the human body. Such representation can be built from different cues (even combined) and carries important advantages in person re-identification and people search tasks, which will be discussed in the next Sections.

In the rest of the Chapter, the reader is first introduced to dissimilarity-based representations in Sect. 3.1. The MCD framework itself is then presented and motivated in Sect. 3.2. Sect. 3.3 and Sect. 3.4 show respectively how MCD representations can be exploited for the tasks of person re-identification and people search. An extension of the framework to combine different cues is also presented in Sect. 3.5. Sect. 3.6 concludes the Chapter, and provides a brief outline of the experimental analysis provided in the next Chapters.

3.1 Dissimilarity-based representations for pattern recognition

Pattern recognition is a field of study devoted to the design of systems able to automatically recognise a particular kind of object or distinguish among categories (*classes*) of objects. The traditional approach to pattern recognition [42] is indeed inspired by the modern scientific method, which builds on empiric observations, measurements of phenomena, and a subsequent formulation of a theory or *model*, that describes them and hopefully allows one to make predictions about them. In particular, pattern recognition usually follows the simple scheme of i) describing objects as sets of measurements, called *features* (e.g., colour, weight, length, etc.), ii) use statistical classifiers to learn a model of the classes of objects of interest in the feature space, from a *training set* of labelled examples, and iii) generalise the learned model to unseen objects described in the same feature space.

Indeed, this approach is rational, simple and works very well in a number of practical cases. However, the success of such approach is directly, strongly influenced by a proper choice of the features used. Unfortunately, in many classification problems it may be difficult to find a suitable set of features. An example is how to describe the shape of an object (see the example on Fig. 1.4).

Dissimilarity representations have been originally proposed to deal with classification problems in which a feature vector representation is not available, or it is difficult to find discriminative features, and it is possible to define a dissimilarity measure between pairs of objects instead. By means of such measure, it is possible to represent any object with a vector of dissimilarity values to a predefined set of "prototype" objects. This novel object representation is then used in place of classic feature vectors to address pattern recognition problems. Dissimilarity-based representations build on the fact that the notion of similarity (or dissimilarity) of objects is more *fundamental* that their description in terms of measurements: indeed, a strong paradigm shift with respect to feature vector representations[1].

Formally, given a set of n objects $X = \{x_i\}$ and a representative set of m prototypes $P = \{p_j\}$, the dissimilarity representation of X is defined by means of a $n \times m$ dissimilarity matrix

$$D = \left(d_{ij}\right) \tag{3.1}$$

where

$$d_{ij} = d\left(x_i, p_j\right) \tag{3.2}$$

is the dissimilarity of the pair of objects x_i and p_j. Each element x_i of X is then represented as the i-th row of D. Note that the prototype set can be a subset of X, or even coincide with X. In the latter case, D is square.

The distance measure $d(\cdot,\cdot)$ is called a *metric* when the following conditions hold:

1. *reflectivity*: $d(x,x) = 0$
2. *positivity*: $d(x,y) > 0$ if $x \neq y$
3. *symmetry*: $d(x,y) = d(y,x)$
4. *triangle inequality*: $d(x,y) < d(x,z) + d(z,y)$ for every z

While *reflectivity* and *positivity* are crucial to define a proper dissimilarity measure $d(\cdot,\cdot)$, the latter two properties are actually not essential [105], which is important given that non-metric distances often seem to arise e.g. in computer vision [41, 71]. This fact also enables the use of dissimilarities derived from psychological judgements, that often lack of the symmetry property [128].

Different approaches can be followed to define the prototype set P. The straightforward way is to simply take the whole data set X as prototype set. However, as the number of samples in X grows, the size of the dissimilarity vector associated to each sample may become too high, and ultimately lead to an high computational complexity and to a reduction of the discriminant capabilities of classifiers in classification tasks (the so-called *curse of dimensionality* [42]). Thus, some prototype selection scheme [103] should be put in place to properly choose the elements of P. In the following, some of the most representative prototype selection schemes are briefly reviewed.

[1] Pekalska and Duin [104] extensively discussed this point in their book on dissimilarity representations [104], which has largely inspired this thesis work. The interested reader is pointed to this book for examining in depth the theoretical motivations on dissimilarity-based representations.

Random selection. $m < n$ prototypes are randomly chosen from X. Alternatively, k prototypes are randomly chosen for each class ω_c of X.

Clustering. Elements of X are clustered into m clusters (e.g. by using k-means [42]) and prototypes are chosen as the elements of X nearest to the each centroid. Note that this may require that a vectorial representation of the elements of X is available, as clustering techniques usually work on vectorial spaces.

Mode seek. Prototypes consist of the modes estimated from each class ω_c in X. For each ω_c, the algorithm proceeds as follows [29]:

1. choose a relative neighbourhood size $s > 1$ (s integer);

2. for each object $x_i \in X_{\omega_c}$, where X_{ω_c} is the subset of X of elements belonging to the class ω_c, find the dissimilarity $d\big(x_i, \mathrm{nn}_s(x_i)\big)$ to its s-th nearest neighbour;

3. find a set P_{ω_c} consisting of all $x_j \in X_{\omega_c}$ for which $d\big(x_j, \mathrm{nn}_s(x_j)\big)$ is minimum within its set of s nearest neighbours.

The objects from the set P_{ω_c} are the estimated modes of the distribution of ω_c in terms of the given dissimilarities. The final prototype set is the union of P_{ω_c}: $P = \bigcup_c P_{\omega_c}$.

Feature selection. First, the prototype set is defined as the whole X. Dissimilarities are then treated as features and a feature selection technique [61] is used to select prototypes whose corresponding dissimilarity values carry useful information for the classification problem at hand.

Editing. Similarly to the case of feature selection, prototypes are at first defined as the whole X. An editing algorithm [42] is then applied to the dissimilarity matrix $D(X, X)$ to reduce the space complexity, e.g. by eliminating prototypes that are surrounded by objects of the same class.

Interestingly, it has been shown [103] that performance of classification algorithms that work in a dissimilarity space may be only slightly affected by the way prototypes are chosen (the prototype selection scheme that usually lead to worst performance is the random selection). The *number* of prototypes that are used, instead, seems more important (an higher number of prototypes usually guarantees a lower classification error).

Dissimilarity representations threat objects in their *wholeness*, i.e., dissimilarities are evaluated between pairs of entire objects. In some pattern recognition tasks, especially in computer vision ones, the objects of interest have a *structure* that is better described with a part-based model. E.g., the human body, which has a complex kinematics and can appear in different images with different poses, is often represented as a collection of parts (see Sect. 2.1). Moreover, there can be *localised* characteristics of objects that need to be preserved in the representation (e.g., local features of the appearance of the human body). For this kind of objects, dissimilarity representations as described above may not be suitable. In the next Section, they will be extended to proper describe objects exhibiting multiple parts and localised characteristics.

3.2 Multiple Component Dissimilarity representations

Many objects may be better described as a rigid or non-rigid collection of parts. The human body, for instance, may be decomposed into different body parts, e.g. the head, the torso, the arms, and the legs, whose relative position may be constrained by specific relations. As

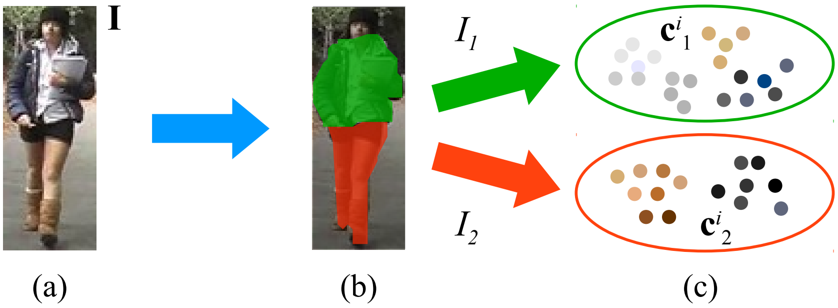


Figure 3.1: General representation of the human appearance adopted by MCD. (a) The image of an individual. (b) The body is subdivided into body parts: in this example, upper and lower body, shown respectively in green and in red. (c) A set of components (e.g., SIFT points) is extracted form each body part. Components are represented here as coloured dots.

stated in Sect. 2.1, this fact has been exploited by various person re-identification methods to construct better appearance descriptors.

Generally speaking, the human body appearance[2] can be described as an ordered sequence of M sets, corresponding to M body parts ($M \geq 1$):

$$\mathbf{I} = \{I_1, \dots, I_M\}. \tag{3.3}$$

Each set I_m may contain a number of *local features*, each represented by a feature vector $\mathbf{c}^k_m$, or one single global feature vector $\mathbf{c}^1_m$ describing the appearance of the whole body part:

$$I_m = \{\mathbf{c}^i_m\}, \qquad \mathbf{c}^i_m \in \mathbb{X}. \tag{3.4}$$

These $\mathbf{c}^i_m$ are called *components* in MCD terminology. $\mathbb{X}$ is the *feature space*, which for the sake of simplicity is assumed to be the same for all the components, without loosing generality. Fig. 3.1 visually describes this model in the case when two body parts are used.

This model for the human appearance is general and can frame all existing appearance descriptors. For instance, the descriptors of [8, 9, 45, 55, 93] extract multiple body parts; the ones of [35, 45, 55, 58, 63] use multiple components; in [45, 93] both a body part subdivision and a multiple component representation is used. Even descriptors made up of one single feature vector (e.g., [20, 126]) can be viewed as a particular case of the multiple parts/multiple components representation, where only one "component" is extracted from one single body "part" (the whole body).

Note that the number of sets of components can be higher than the number of body parts, if more than one kind of features is used to describe each part. An example is the popular SDALF descriptor [45], which in Fig. 3.2 is shown as an instantiation of the multiple parts/multiple components model. SDALF uses a two body-part subdivision into *torso* and *legs*, discarding the image region corresponding to the head. From each body part, *Maximally Stable Colour Regions* (MSCR) and *Recurrent High-Structured Patches* (RHSP) are extracted. Finally, a weighted HSV histogram is extracted from each body part. The two

[2] The rest of the Chapter will refer specifically to the representation of the human body appearance, as this is the object of interest of person re-identification. Note however, that the same multiple parts-localised characteristics description is suitable for a vast number of objects, e.g. non-rigid objects like animals and moving mechanisms, or rigid ones like cars.

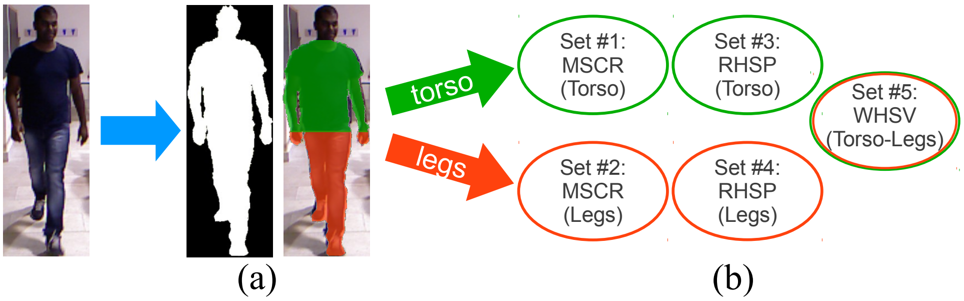


Figure 3.2: The popular SDALF descriptor [45] as an instantiation of the proposed appearance model. SDALF subdivides body into torso (disregarding the head) and legs parts (a); from each part, RHSP and MSCR local features are extracted, and the weighted HSV histograms of each part are concatenated, leading to five sets of components (b).

weighted histograms are then concatenated to form a single feature vector. The total number of sets of components is therefore 5.

Consider now a *gallery* $\mathscr{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ of N individuals, each described as above. The goal is to build a dissimilarity-based representation of such individuals, preserving their multiple parts/multiple components structure. To this aim, a set of K_m prototypes $P_m = \{\mathbf{p}_m^k\}$ is first defined for each body part. Prototypes are components chosen from $\mathscr{I}$ as described below. For each $\mathbf{I} \in \mathscr{I}$, a dissimilarity descriptor $\mathbf{I}^{\mathrm{D}}$ is then created, as a concatenation of the vectors of dissimilarity values between each $I_m \in \mathbf{I}$, and the prototypes P_m of the m-th body part.

Prototypes are created as follows. For each body part $m = 1, \ldots, M$:

1. The feature vectors of the m-th part of each $\mathbf{I} \in \mathscr{I}$ are merged into a set $X_m = \bigcup_{j=1}^{N} I_{j,m}$;

2. K_m prototypes of the m-th body part are chosen from X_m to form the prototype gallery $P_m = \{\mathbf{p}_m^1, \ldots, \mathbf{p}_m^{K_m}\}$.

Step 2 can be carried out with any prototype selection technique like those described in Sect. 3.1. Note that, regardless of the prototype selection method adopted, prototypes encode low-level local or global visual characteristics of the appearance.

The above procedure returns M sets of prototypes, one for each body part:

$$\mathbf{P} = \{P_1, \ldots, P_M\} \,. \tag{3.5}$$

Fig. 3.3 visualises the procedure of prototype selection when k-means is used with a descriptor having two body parts.

Once the prototypes have been defined, given the original multiple parts/multiple components descriptor of any individual, $\mathbf{I} = \{I_1, \ldots, I_M\}$, its MCD descriptor is obtained as the concatenation of the M dissimilarity vectors

$$\mathbf{I}^{\mathrm{D}} = \left[I_1^{\mathrm{D}} \quad \ldots \quad I_M^{\mathrm{D}} \right], \tag{3.6}$$

where:

$$I_m^{\mathrm{D}} = \left[d\left(I_m, \mathbf{p}_m^1\right) \quad \ldots \quad d\left(I_m, \mathbf{p}_m^{K_m}\right) \right], \quad m = 1, \ldots, M \,, \tag{3.7}$$

and $d\left(I_m, \mathbf{p}_m^j\right)$ is the dissimilarity between I_m and the j-th prototype of the m-th body part. Since I_m is a set of components, and p_m^j is a single component, the dissimilarities $d\left(I_m, \mathbf{p}_m^j\right)$

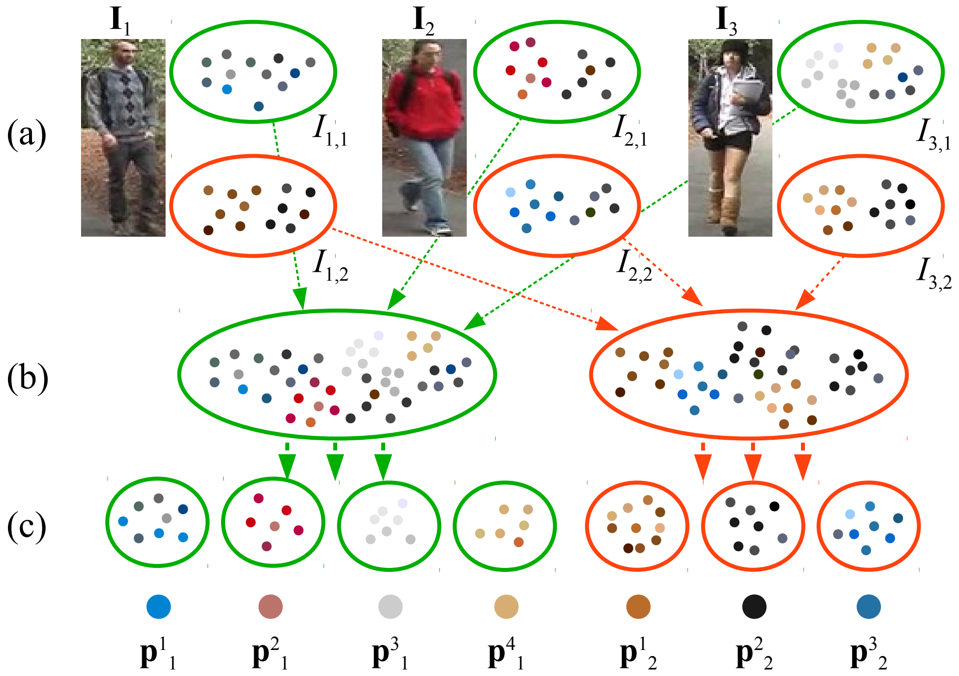


Figure 3.3: Creation of the prototype gallery in MCD using k-means for prototype selection. In this example, the body is subdivided into two parts: upper (in green) and lower body (in red). (a) A gallery of three individuals, represented according to MCM. (b) All the components of the same part are merged. (c) The k-means clustering algorithm is applied (in this example, with $k = 4$ for the upper body part and $k = 3$ for the lower body part), and prototypes are selected for each part as the components nearest to the centroids of each cluster.

must be evaluated via an one-vs-many distance measure, e.g. the minimum of the distances of between each element of I_m and $\mathbf{p}^j_m$.

This dissimilarity-based representation exhibits a considerable reduction in storage requirement: only one vector of real values for each individual, and the set of prototypes, need to be stored. The compactness of dissimilarity descriptors also allows for extremely fast matching between pairs of them, as computing distances between vectors of real numbers is a very fast operation with modern CPUs. This can enable several useful applications, like real-time re-identification of an individual, among a huge number of candidates. Moreover, prototypes represent localised low-level characteristics that can encode high-level concepts, which as will be shown in Sect. 3.4 can enable appearance-based people search over images described by dissimilarity vectors.

It must be pointed out that a seemingly analogous representation is used also in *visual words* methods, widely used e.g. in scene categorization (e.g., [135]). In these methods, a visual codebook is first built off-line, then the *frequency* (count of the occurrences) of each visual word *inside* each sample is considered. Differently, in the dissimilarity paradigm the *degree* of similarity of each prototype to the *whole* sample is considered.

3.3 Multiple Component Dissimilarity matching for person re-identification

Person re-identification is usually modelled as a *matching* problem (see Sect. 1.1), whose goal is to rank templates in a given *template gallery* with respect to their similarity to a given *probe* individual. The above dissimilarity-based representation can be conveniently used to match appearance descriptors in the dissimilarity space.

Let $\mathscr{T}$ be a template gallery of N individuals, and $\mathbf{Q}$ be a probe individual. Their MCD representation is denoted respectively as $\mathscr{T}^{\mathrm{D}} = \{\mathbf{T}_1^{\mathrm{D}}, \ldots, \mathbf{T}_N^{\mathrm{D}}\}$ and $\mathbf{Q}^{\mathrm{D}}$. Note that the same sets of prototypes must be used for constructing the MCD representation of all the templates and of the probe. Such sets can in principle be defined from the any gallery of individuals, including the template gallery, or a different design data set. Similarly to Eq. (1.1), the problem of re-identifying the individual $\mathbf{Q}^{\mathrm{D}}$ in the dissimilarity space can be formulated as:

$$\mathbf{T}^{*\mathrm{D}} = \arg\min_{\mathbf{T}_i^{\mathrm{D}}} D^{\mathrm{D}}\left(\mathbf{T}_i^{\mathrm{D}}, \mathbf{Q}^{\mathrm{D}}\right) . \tag{3.8}$$

Where $D^{\mathrm{D}}(\cdot,\cdot)$ is a distance measure in the dissimilarity space. It plays an important role, as a suitable distance measure can lead to a better performance. In principle, any common distance measure, like the Euclidean, cosine, and normalised cross-correlation distances, could be used. However, none of the above measures properly captures the concept that underlies the proposed dissimilarity representation, that is, each dissimilarity value represents a degree of *presence* (and then, of *relevance*) of the corresponding prototype. Thus, every element of a dissimilarity vector carries a different amount of information in representing the sample of interest. In particular, lower dissimilarity values carry more information than higher values, and thus encode the most relevant characteristics of the sample.

Based on the above arguments, here a weighted Euclidean distance between a pair of dissimilarity vectors $\mathbf{x}$ and $\mathbf{y}$ associated to two given objects is proposed. Each weight reflects the importance of the corresponding prototype with respect to such objects:

$$d^{\mathrm{D}}(x,y) = \left(\sum_i \frac{w_i}{W}|x_i - y_i|^2\right)^{1/2}, \tag{3.9}$$

where W is a normalization factor such that $\frac{1}{W}\sum_i w_i = 1$. The weights w_i must be chosen to guarantee that prototypes carrying more information about at least one of the objects receive higher relevance (i.e. an high w_i). On the other hand, the weight of prototypes that are less important for both objects of them must be low. In other words, when comparing two objects one must look mostly at visual characteristics that are *present* in at least one of the two. To do so, the weights w_i are defined as

$$w_i = \mathrm{f}(\overline{w}_i), \tag{3.10}$$

$$\overline{w}_i = 1 - \min(x_i, y_i), \tag{3.11}$$

where $\mathrm{f}(\cdot)$ is a monotonically increasing function, and $\overline{w}_i$ is the maximum similarity (corresponding to the minimum dissimilarity) of the i-th prototype with respect to both objects, assuming that dissimilarity values x_i and y_i are in the range $[0,1]$. Thus, the higher the relevance (the lower the dissimilarity) of the i-th prototype to at least one of the objects, the higher the w_i.

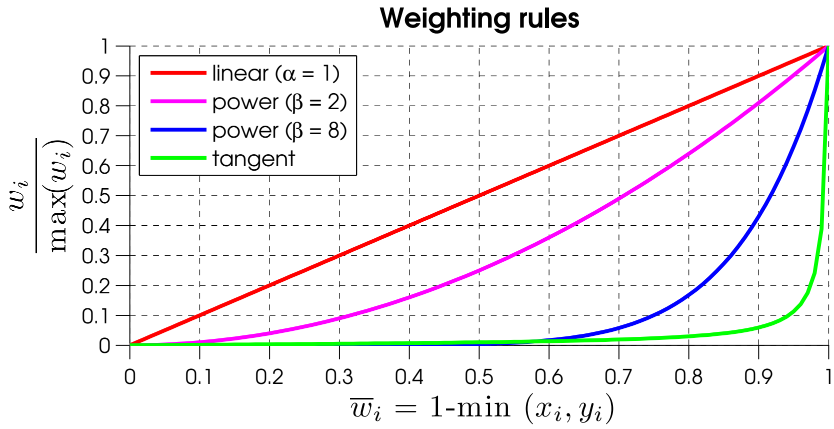


Figure 3.4: Comparison of weighting rules for the weighted Euclidean distance in the dissimilarity space, normalised to one. Note that for the Tangent rule $\overline{w}_i$ has been truncated to 0.99 in order to avoid Infinite weights.

Different choices of the weighting rule f(·) can enhance the differentiation of relevant prototypes from non-relevant ones. Here three possible choices are proposed:

- **linear:** $f(\overline{w}_i) = \alpha\overline{w}_i$;
- **power:** $f(\overline{w}_i) = \overline{w}_i{}^{\beta}$;
- **tangent:** $f(\overline{w}_i) = \tan\left(\frac{\pi}{2}(\overline{w}_i\right)$.

The linear rule shall work well when the degree of presence of a prototype varies pretty much linearly with the corresponding dissimilarity value. The power and the tangent rules, instead, strongly differentiate between high and low dissimilarities, and shall be used when the relevance of certain prototype is assumed to be high only if the corresponding dissimilarity is very low. Fig. 3.4 shows the weight, normalised to one, as a function of $\overline{w}_i$ using the three rules above.

3.4 Multiple Component Dissimilarity and people search

The Multiple Component Dissimilarity representation can be conveniently used also for a novel task, consisting of finding, among a set of images of individuals, the ones relevant to a *textual* query that describes the clothing appearance of an individual of interest. This novel task is named *appearance-based people search* in this thesis, and differs from person re-identification, where the query is an image of the person of interest. This can be useful in applications like forensics video analysis, where a textual description of the individual of interest given by a witness can be available, instead of an image.

In the following, a general approach to extend based person re-identification systems to enable also the people search functionality is proposed, based on MCD representations. The approach is based on the intuition that the high level concepts that form the textual query and describe a certain clothing characteristics (e.g., “red shirt”), may be encoded by one or more visual prototypes, according to the low-level features and part subdivision used. As

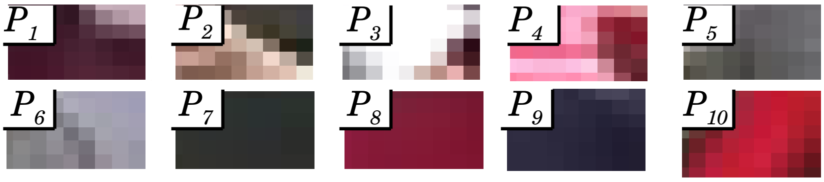


Figure 3.5: Prototypes obtained from the upper body parts of a small set of individuals. Descriptors of people wearing a red shirt should exhibit a high similarity to prototypes $\mathbf{p}_8$ and $\mathbf{p}_{10}$. A high similarity to $\mathbf{p}_3$ can be expected instead in the case of a white shirt.

an example, consider the 10 prototypes shown in Fig. 3.5, selected using k-means clustering from a set of rectangular patches, represented by HSV histograms, sampled from the upper body parts of 24 images of individuals from the VIPeR data set [57]. Intuitively, descriptors of people wearing a red shirt should exhibit a high similarity to prototypes $\mathbf{p}_8$ and $\mathbf{p}_{10}$, while a high similarity to $\mathbf{p}_3$ can be expected in the case of a white shirt. Similarly, other prototypes may encode useful information to recognise other clothing characteristics.

Following the above intuition, a possible approach to perform appearance-based people search through an existing appearance descriptor, is to:

(i) identify a set $\mathscr{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \ldots\}$ of clothing characteristics that can be detected by the given descriptor, named *basic queries*;

(ii) construct a detector for each basic query $\mathbf{Q}_i$, using dissimilarity values as *features* of a supervised classification problem.

The basic queries that have to be identified in step (i) depend on the original descriptor. For instance, if it separates lower and upper body parts, and uses colour features, one basic query can be "red trousers/skirt". Step (ii) can be viewed as a supervised binary classification problem for each $\mathbf{Q}_i$, which consists of recognising the presence or absence of the corresponding visual characteristic, using as features the dissimilarity values between an image descriptor and the prototypes. A binary classifier (e.g., a Support Vector Machine [31]) can be trained using as features the dissimilarity values of an image descriptor to the prototypes. The training set can be obtained from a gallery of images of individuals, labelled accordingly. The resulting classifier can then be used as the detector for the basic query $\mathbf{Q}_i$.

Note that one may know in advance that some features (prototypes) do not carry any discriminant information for some $\mathbf{Q}_i$. For instance, this is the case of the prototypes of the lower body part, with respect to queries related to the upper body. Such features can thus be discarded before constructing the corresponding classifier.

Finally, complex queries can be built by connecting basic ones through Boolean operators, e.g., "red shirt AND (blue trousers OR black trousers)". Given a set of images, those relevant to a complex query can simply be found by combining the subsets of images found by each basic detector, using the set operators corresponding to the Boolean ones. In the above example, this amounts to the union (OR) of the images retrieved by the "blue trousers" and "black trousers" basic queries, followed by the intersection (AND) with the images retrieved by the basic query "red shirt".

The above approach for building detectors is independent of the original appearance descriptor, as it is based on the general framework of MCD.

3.5 Combination of multiple modalities

The clothing appearance is used in the most of re-identification methods as a soft, session-based cue, that is relatively easy to extract, and exhibits uniqueness over limited periods of time. However, the recognition performance that can be attained by using clothing appearance is limited, especially in scenarios where the number of individuals is very large, since many individuals could wear very similar clothing. In a recent paper [28] recognition performance of a proposed appearance-based re-identification method was compared to the one of human operators, to assess how far computer vision is from the empirical "upper bound" represented by human performance. Although human operators outperform the machine, as expected, they surprisingly achieve an average first-rank recognition rate of 75%, which is lower than one could expect, given that the individual in each query image had to be found among a set of 45 pedestrian images only. This suggests that clothing appearance can be an intrinsically "weak" cue for re-identification,[3] and that combining it with other biometric traits may be useful. The extraction of other soft biometrics can be enabled by the recent introduction of RGB-D sensors, like MS Kinect, even in unconstrained environments and without calibration, thanks to their per-pixel depth estimation that is added to the usual RGB information. For instance, human pose estimation capability of Kinect SDK [80] provides 20 different skeletal points in metric coordinates, that can be used to estimate various anthropometric measures (e.g. the height, the arm length, and others [10]).

When different biometric modalities (like clothing appearance and anthropometric measurements) are used in matching problems like person re-identification, a proper fusion strategy must be used to combine the respective information. In principle, two main fusion techniques can be exploited to this aim [114]:[4]

1. *feature*-level fusion: feature vectors coming from different modalities are concatenated into a single feature vector;
2. *score*-level fusion: a distinct detector/matcher is used for each modality, and their real-valued scores are combined (e.g., by averaging them, or using their maximum value).

Feature-level fusion can be applied if each modality is represented by a fixed size feature vector, and features exhibit an intrinsic ordering. However, this is not the case for most appearance-based descriptors used in re-identification, may be made up of multiple local features, without an intrinsic ordering between them (see Sect. 2.2). Other modalities may have a different representations, e.g. anthropometric measures are typically scalar values (see Sect.2.4). Even when concatenating feature vectors is possible, the resulting size could arise computational and *curse of dimensionality* issues [18]. On the other hand, the performance of score-level fusion can be strongly affected by the choice of the fusion rule and of its parameters (if any), and a suitable choice for the task at hand may be not trivial.

[3] It is easy to see that the human could reason only on the clothing appearance to perform the task, if one looks at the images used in the test. In fact, the face was often not recognizable, because of the low resolution of the images and of the different poses. Moreover, the images showed only the region of the original frame containing the person, so that it was difficult to estimate body measurements like the height, that could have helped in distinguishing the target individual from the others.

[4] In *verification* tasks, whose goal is to establish whether the claimed identity is true, multi-modal fusion can also be performed at *decision* level, i.e., by combining the crisp outputs of classifier/detectors. It can not be applied to person re-identification, which is a *recognition* task instead.

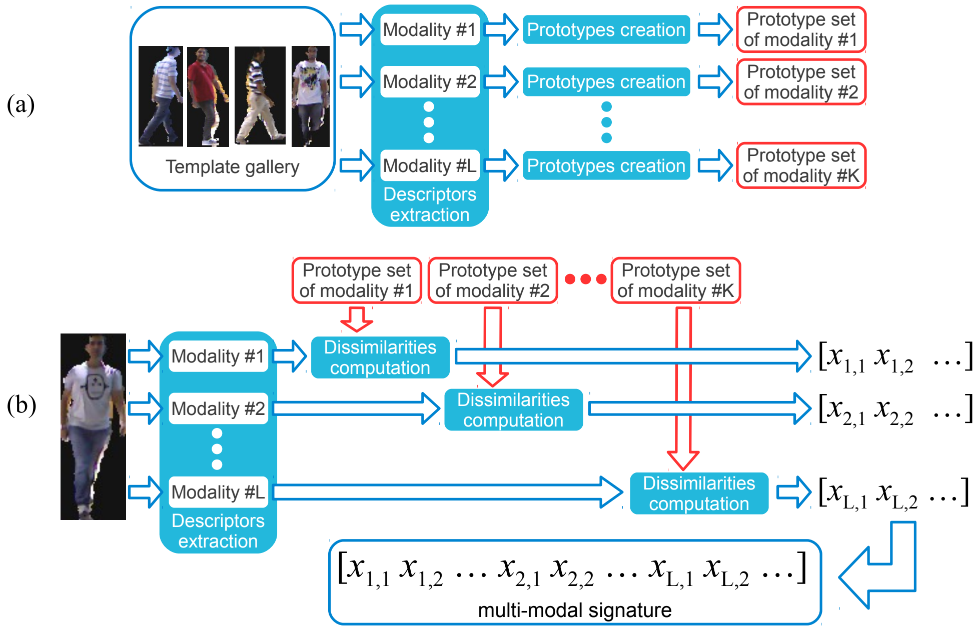

(a)
Template gallery
Modality #1
Modality #2
Modality #L
Descriptors extraction
Prototypes creation
Prototypes creation
Prototypes creation
Prototype set of modality #1
Prototype set of modality #2
Prototype set of modality #K
(b)
Prototype set of modality #1
Prototype set of modality #2
Prototype set of modality #K
Modality #1
Modality #2
Modality #L
Descriptors extraction
Dissimilarities computation
Dissimilarities computation
Dissimilarities computation
$[x_{1,1}\ x_{1,2}\ \ldots]$
$[x_{2,1}\ x_{2,2}\ \ldots]$
$[x_{L,1}\ x_{L,2}\ \ldots]$
$[x_{1,1}\ x_{1,2}\ \ldots\ x_{2,1}\ x_{2,2}\ \ldots\ x_{L,1}\ x_{L,2}\ \ldots]$
multi-modal signature

Figure 3.6: Scheme of the extended MCD framework. (a) *prototype construction*. First, multiple parts/multiple component descriptors in each modality are extracted from each individual in the template gallery. Then, a separate set of prototypes for each modality is constructed. (b) *multi-modal descriptor computation*: each template individual, and the probe individual, are described in the dissimilarity spaces defined by the sets of prototypes of each modality. The dissimilarity vectors obtained are finally concatenated to obtain the multi-modal dissimilarity descriptor.

The MCD framework can provide a third way to combine different modalities in an uniform and elegant fashion. In fact, an important characteristic of MCD descriptors is that they are *representation-independent*, in the sense that the underlying representation (e.g., feature vectors, if available) of prototypes and objects, is actually not relevant from the point of view of the dissimilarity representation. Prototypes are logically and semantically at an higher level than the actual features extracted from objects. This suggested us another way to build multi-modal descriptors: in fact, prototypes can even represent characteristics seen in different modalities. Therefore, one can conveniently create different sets of prototypes for each modality, and combine dissimilarities to these prototypes in a single dissimilarity vector, which will be semantically as coherent as a dissimilarity vector with respect to prototypes of a single modality.

In the following, the MCD framework is extended to support multiple modalities, e.g., clothing appearance and anthropometric measurements. As stated in Sect. 3.2, when clothing appearance is the sole modality used, a person is described as an ordered sequence of sets of components, where each set is associated to a different body part. If L different modalities are used, each one can be associated to a distinct feature vector $\mathbf{v}_l$, $k = 1, \ldots, L$. These feature vectors can be framed into the multiple parts/multiple components representation, by considering each $\mathbf{v}_k$ as a vector of observations coming from one single body part, corresponding to the whole body. In the multiple parts/multiple components representation, this corresponds to the particular case where only one "component" is extracted from the whole body. This enables the application of the MCD framework. In particular, a set of prototypes for each modality has to be constructed first. Then, for any given individual $\mathbf{I}$, L dissimilarity vectors $\mathbf{I}^{\mathrm{D},1},\ldots,\mathbf{I}^{\mathrm{D},L}$ are constructed, corresponding to the considered modalities. To this aim, an appropriate dissimilarity measure between $\mathbf{I}$ and a prototype must be defined, for each modality. The dissimilarity vectors $\mathbf{I}^{\mathrm{D},1},\ldots,\mathbf{I}^{\mathrm{D},L}$ can then be concatenated into a multi-modal dissimilarity vector representation of individual $\mathbf{I}$:

$$\mathbf{I}^{\mathrm{D},multi-modal} = [\mathbf{I}^{\mathrm{D},1},\ldots,\mathbf{I}^{\mathrm{D},L}] \; . \tag{3.12}$$

The proposed multi-modal representation is summarised in Fig. 3.6. It has the advantage of being compact, and feature-independent. Furthermore, two multi-modal dissimilarity vectors can be matched using the same weighted Euclidean distance of Eqs. (3.9)-(3.10). This means in particular that it is not necessary to define an appropriate set of weights for combining the information coming from each modality during matching, since higher weights are automatically given to the most distinctive characteristics (prototypes), regardless of the modality they are associated to.

This extension of MCD representations exhibits important advantages with respect to feature-level fusion rules. First, it enables the combination of modes characterized by fixed-length feature vectors with modes characterized by unordered, multiple local features (e.g. appearance descriptors). Second, it can overcome the dimensionality issue, as dissimilarities can be evaluated with respect to a small number of prototypes, even if the underlying vectorial representation (if any) is made up of big feature vectors. Moreover, it does not need to empirically choose a proper fusion rule (the actual choice of which may strongly affect performance), as per score-level fusion.

3.6 Outline of the following Chapters

In this Chapter, the Multiple Component Dissimilarity framework for describing objects with multiple parts and multiple components (such as the human body) in a dissimilarity space has been presented. The application of MCD to the tasks of person re-identification and people search has been also discussed. The following three Chapters experimentally evaluate the application of MCD to these tasks. Chapter 4 applies MCD to the task of speeding up an existing person re-identification method. Chapter 5 then present a novel appearance-based re-identification method attaining state-of-the-art performance with low computational requirements, where different kinds of global and local features are extracted and combined into a single dissimilarity vector using MCD. Chapter 6 implements the method described in Sect. 3.5 to perform multi-modal person re-identification on networks of RGB-D cameras, combining appearance and anthropometry. Finally, Chapter 7 shows and experimentally evaluates a practical implementation of appearance-based people search.

Chapter 4

Using MCD to speed up existing re-identification methods

Despite the practical relevance of computational complexity of person re-identification methods, this issue has been almost overlooked in the literature so far. Moreover, it is worth pointing out that the processing time of many existing methods may be too high for practical applications. Dissimilarity-based descriptors obtained through the use of MCD have a great advantage with respect to processing time. In fact, they are basically vectors of real numbers, and therefore they are much more compact than, e.g., bags of multiple feature vectors coming from different interest points. Such compactness can drastically reduce their matching time, which is the most time-consuming step of a person re-identification method.

In this Chapter, MCD is exploited for speeding up an existing person re-identification method. In particular, MCD is applied to a baseline method that is representative of a typical appearance descriptor for person re-identification, which extracts random partly overlapping rectangular patches from the body torso and legs separately.

As will be shown in the experimental results reported in this Chapter, this may happen at the expense of a lower re-identification accuracy. Despite this, trading a lower accuracy for a lower processing time can be advantageous in practical applications, and this will be demonstrated through the use of a simple quantitative model.

The rest of the Chapter is structured as follows. First, the model for evaluating the trade-off between accuracy and processing time is presented in Sect. 4.1. A description of the baseline method and details on the application of MCD are provided in Sect. 4.2. A thorough experimental evaluation which analyses both performance and processing time an how they are influenced by the main parameters of MCD is then presented, in Sect. 4.3. Conclusions are finally drawn in Sect. 4.4.

4.1 Trade-off between re-identification accuracy and matching time

MCD-based methods may attain a much lower matching time and memory requirement than its non-dissimilarity-based version. However, this is sometimes attained at the expense of a lower re-identification accuracy, as will be shown in Sect. 4.3. In such a case, re-identification methods obtained by MCD are not advantageous over their non-dissimilarity-based counterparts, in application scenarios where a higher accuracy is more important than a lower processing time. For instance, this can be the case of off-line forensics investigations (e.g., looking for an individual of interest among a dataset of videos previously recorded).

However, trading a lower accuracy for a lower processing time can be advantageous in other scenarios. As an example, consider a real time application in which individuals observed by different, non-overlapping cameras are automatically tracked, and a human operator can select an individual of interest from one of the videos, and ask the system to re-identify it (again, in real time). In this case, the template gallery containing the descriptors of all tracked individuals can be automatically constructed and updated in real time. When the operator sends a probe image to the system, it first builds the corresponding descriptor, then matches such descriptor against all the ones in the template gallery, and returns to the operator the list of templates ranked for decreasing similarity with the probe. Finally, the operator scrolls such list to search for the individual of interest (see Fig. 4.1). Clearly, a re-identification method A with a lower accuracy than another method B results in a higher average search time spent by the operator to find the individual of interest in the ranked list provided by the system (assuming the operator has a 100% accuracy). However, if the higher search time of method A is balanced by a lower processing time, the overall re-identification time between submitting the probe and finding the corresponding individual in the ranked list (namely, the sum of the processing and search times) can be lower for method A than for B. Therefore, in a real time scenario like the one considered above, method A can be preferable to B, although its re-identification accuracy is lower.

In the following a simple quantitative model is given to evaluate the overall re-identification time, that will be used in the experimental evaluation of Sect. 4.3. Let

- t_d be the average time required to construct the descriptor of the image of an individual, using a given method, and
- t_m the average matching time between two descriptors.

If the template gallery contains the descriptors of N individuals, the average processing time t_p is given by the time needed for constructing the probe descriptor plus the time needed to match it to the N template descriptors (as explained above, here it is assumed that template descriptors are constructed during tracking, and are thus already available at this time):

$$t_\mathrm{p} = t_\mathrm{d} + N t_\mathrm{m} \, . \tag{4.1}$$

Let us now denote with t_c the average time spent by the operator to compare the probe image with a template image, and with $R \in \{1, \ldots, N\}$ the random variable defined as the rank of the query individual in the list provided by the system.[1] The average search time t_s spent by the operator is given by:

$$t_\mathrm{s} = t_\mathrm{c} E\{R\} \, , \tag{4.2}$$

where $E\{R\}$ is the expected rank of the probe individual. The value of $E\{R\}$ can be computed from the Cumulative Matching Characteristics (CMC) curve, which is a widely used measure

[1] Here it is assumed that the template gallery always contains the correct match. If this is not the case, for the purposes of this section the domain of R can be extended by adding a value $N+1$, to denote the absence of the correct match in the template gallery.

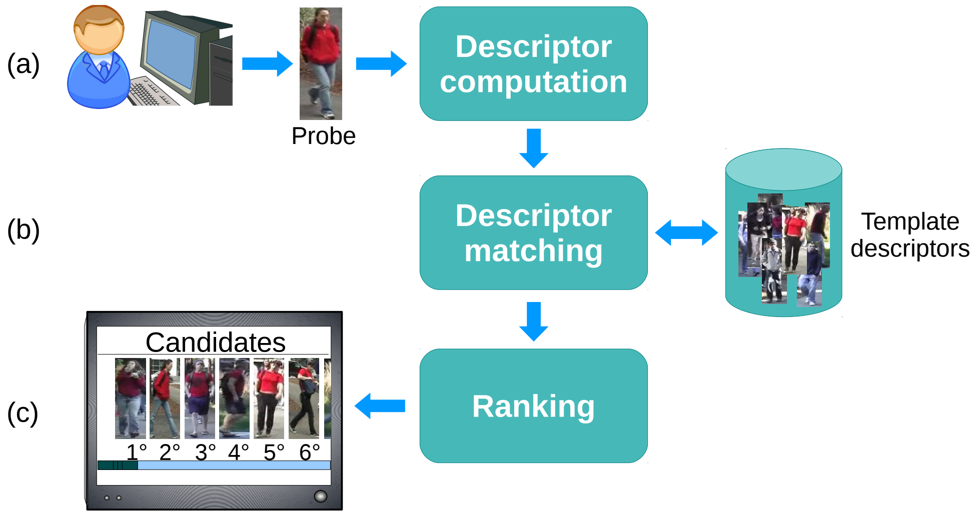


Figure 4.1: A typical on-line, real-time application scenario of person re-identification. a) The security operator sends a probe image to the system. The system builds a descriptor of the probe, and b) matches it against all the descriptors stored in the template gallery, that are constructed on-line. c) The system returns to the operator the list of templates ranked for decreasing similarity to the probe.

of ranking accuracy of re-identification methods. It is defined as the cumulative distribution of R: $P(R \leq r)$, $r = 1, \ldots, N$, namely the probability that the template image of the query individual is among the top-r ranked images. Using the standard notation $CMC(r)$ for $P(R \leq r)$, it is easy to see that $E\{R\}$ is given by:

$$\begin{aligned} E\{R\} = & \sum_{r=1}^{N} rP(R=r) \\ = & \; CMC(1) + \sum_{r=2}^{N} r\big(CMC(r) - CMC(r-1)\big) \\ = & \; N \cdot CMC(N) - \sum_{r=1}^{N-1} CMC(r) \; . \end{aligned} \tag{4.3}$$

The overall average re-identification time t_r can be finally obtained as:

$$\begin{aligned} t_\mathrm{r} = t_\mathrm{p} + t_\mathrm{s} = \; & t_\mathrm{d} + N t_\mathrm{m} + \\ & t_\mathrm{c} \left[N \cdot CMC(N) - \sum_{r=1}^{N-1} CMC(r) \right] \; . \end{aligned} \tag{4.4}$$

This model can be used to compare re-identification methods for the on-line task of Fig. 4.1. In this scenario, the best method is usually the one with the lowest t_r.

4.2 Baseline method and application of MCD

In this Section, MCD is applied to a baseline person re-identification method to lower its average re-identification time t_r. Such baseline method is a straightforward implementation of the multiple parts and multiple components model that underlies MCD. It has been published by the Author in [116], and attains a recognition performance similar to other state-of-the-art methods such as the widely cited SDALF [45]. The method is referred to as “MCMimpl” from here on, which stands for “Multiple Component Method implementation”. It is described in the following.

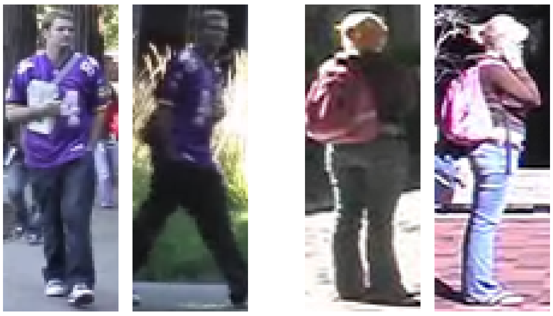


Figure 4.2: Two examples of different images for the same individual: note the difference both in contrast and brightness.

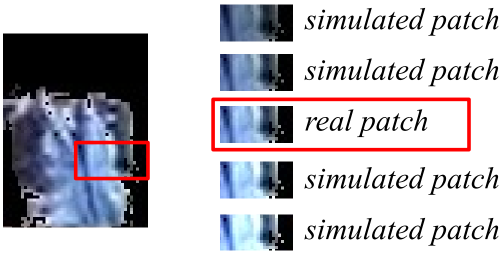


Figure 4.3: Examples of four artificial patches simulating changing illumination (right), corresponding to the patch highlighted on the left.

Given an image of an individual, first background and foreground are separated through a STEL generative model [76]. Then the body is divided into $M = 2$ parts, torso and legs, exploiting its anti-symmetry properties, via the algorithm proposed by Farenzena et al. [45] and described in Sect. 2.1.2. Similarly to [45], the head is discarded, since it does not carry enough information due to its small size. From each part, a set of 80 partly overlapping patches is randomly extracted and represented via the concatenation of H, S and V colour histograms (24, 12 and 4 bins respectively). The patch width and height are defined as 15% of the width and height of the corresponding part.

To increase robustness to illumination changes, these are *simulated* by constructing artificial patches from real template ones.

Light variations usually result in a change of both brightness and contrast of the image (see for example Fig. 4.2). Brightness variations can be obtained by adding or subtracting a fixed value to the RGB components of the pixels of the image. Instead, changing contrast means increasing or decreasing the differences between pixel values. A standard method to obtain this is the following: denoting as $[0, C]$ the original range of each colour channel (usually $C = 255$), every R, G, and B pixel value is translated to $[-C/2, C/2]$, multiplied by a fixed coefficient, and then re-normalised to $[0, C]$. A coefficient greater than 1 results in a higher contrast, while a lower contrast is obtained by choosing values smaller than 1.

To change both brightness and contrast, here a modification of the above technique is used, which does not translate values to $[-C/2, C/2]$ first, but simply multiplies each pixel value of each channel by a coefficient K. Intuitively, this increases (or decreases) the differences between pixel values as well. However, while in the standard method values lower than $C/2$ are reduced, and those higher than $C/2$ are increased, in our variant all the values are increased (or decreased), thus obtaining also a change of brightness. The proposed algorithm multiplies pixel values by a series of coefficients $[k_1,\dots,k_S]$ to generate S simulated patches from each real one (see the example in Fig. 4.2). To choose proper k_i values, the algorithm starts from an initial vector $K=[k_1,\dots,k_S]$, then decreases its values until applying the greatest k_i to the original image does not saturate the image too much. More precisely, it checks that the mean value of R, G and B multiplied by the greatest value of K is not higher than a threshold, which has been set to 240.

The distance between two sets of patches X and Y corresponding to the same part of two different individuals is evaluated by the k-th Hausdorff set distance proposed by Wang and Zucker [132]:

$$d_H(X,Y)=\max(h(X,Y),h(Y,X)) \tag{4.5}$$

where

$$h_k(X,Y)=\underset{x\in X}{kth}\min_{y\in Y}(\|x-y\|) \tag{4.6}$$

where the parameter k governs the sensitiveness to outlying matches, and was set to $k=10$. The the norm $\|x-y\|$ in Eq. (4.6) is the Bhattacharyya distance between histograms, which is defined as:

$$\|x-y\|=\sqrt{1-\sum_i\sqrt{x_iy_i}} \tag{4.7}$$

The final matching score between a probe $\mathbf{Q}$ and a template $\mathbf{T}$ is computed as the average of the distances between the two parts:

$$D(\mathbf{T},\mathbf{Q})=\frac{1}{2}\big(d(T_1,Q_1)+d(T_2,Q_2)\big)\,. \tag{4.8}$$

4.2.1 Prototype creation

To apply MCD to the MCMimpl method described above, prototypes are at first chosen from the template gallery via a two-stage clustering scheme. The Mean-Shift algorithm [30] is used at the first stage, to separately cluster the patches of each individual (excluding the simulated ones), while k-means is applied at the second stage on the resulting centroids.

To incorporate the simulation algorithm of the baseline method, each prototype was finally associated to a set containing 1) the patch nearest to each centroid, and 2) the series of simulated patches created from that patch. A prototype is therefore defined as a set of components: the patch found using the clustering scheme mentioned above, and the set of simulated ones.

The bandwidth parameter of Mean-Shift, which governs the spread of each cluster, was set to $BW=0.3$. The number K_m of prototypes for the m-th body part corresponds to the k value of the k-means algorithm, and therefore it must be set in advance. Although this seems a drawback (as in practice it is difficult to guess a suitable value for K_m), the following experiments will show that the choice of k is not crucial. Mean-Shift, which does not require to define the desired number of clusters beforehand, turned out to produce too unbalanced clusters at the second stage instead, as many of them were composed by only one or two components.

4.2.2 Computation of dissimilarities and matching

The simulated patches added to prototypes make them *sets of components* instead of single components. This means that dissimilarities have to be computed with a set distance, as also the original descriptor is made up of sets of components. A suitable measure can be the k-th Hausdorff distance of Eqs. (4.5)-(4.6). Using the same Bhattacharyya distance of Eq. (4.7) as a metric between pairs of components, which is bounded in $[0,1]$, it is guaranteed that dissimilarities fall in the same range. Thus, during matching the weighted Euclidean distance of Eqs. (3.9) - (3.10) can be used. Among the weighting rules listed in Sect. 3.3, in the following experiments the Tangent rule is used.

4.3 Experimental evaluation

In this Section, MCMimpl is compared with its dissimilarity-based version (denoted in the following as MCMimpl$^{\text{Dis}}$), via a thorough experimental evaluation. The set-up of the experiments is described in Sect. 4.3.1. Experimental results are then provided in Sect. 4.3.2.

4.3.1 Experimental set-up

The classical experimental set-up for assessing person re-identification methods, which is usedin this Section, uses two galleries of images of people, the *template* and the *probe* gallery. Each image is associated to one identity, and each identity is represented by at least one template image and one probe image. For each image of the probe gallery, the templates are ranked with respect to their similarity, and the performance of the algorithm is usually evaluated as the quality of the ranking (the higher is the rank of the true template identity, the better the ranking), e.g. by using the CMC curve defined in Sect. 4.1.

Two benchmark datasets used in many previous works have been used in this experimental analysis: VIPeR [57], and a set of images taken from the i-LIDS MCTS video dataset [139].

The VIPeR dataset is made up of two non-overlapping views of 632 different pedestrians, taken from two different cameras, under different poses, viewpoint and lighting conditions (see Fig. 4.4). It is the most challenging dataset currently available for person re-identification. The first and second view of each pedestrian were used respectively as the template gallery and the probe gallery. The experiments have been carried out on three different subdivisions of this dataset. One of them was used in many previous works: the images of the 632 pedestrians are split into ten, partially overlapping folds of 316 individuals, to carry out ten different runs of the experiments. The same folds have been defined in [45] and are used in this work to obtain results that are comparable with the rest of the literature. Since a template gallery of 316 individuals is relatively small for some real applications, two further subdivisions of VIPeR have been considered: ten, partially overlapping folds of 474 individuals randomly sampled from the whole dataset, and a single fold made up of all the 632 available individuals. This also allowed us to evaluated the trade-off between accuracy and re-identification time as a function of the number of templates. In the following, the above three versions of VIPeR are referred to respectively as VIPeR-316, VIPeR-474, and VIPeR-632. Note that in VIPeR-632 only one run of the experiments is carried out.

The i-LIDS dataset contains 476 images of 119 different pedestrians, taken at an airport arrival hall from different non-overlapping cameras. It shows pose and lighting variations,

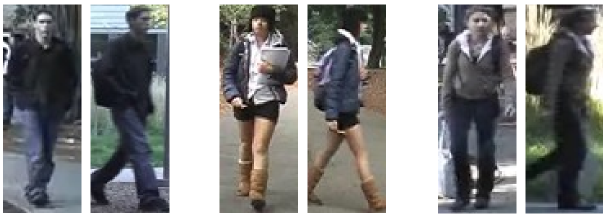


Figure 4.4: Pairs of images showing the same individual taken from the VIPeR data set. Notice pose and illumination variations.

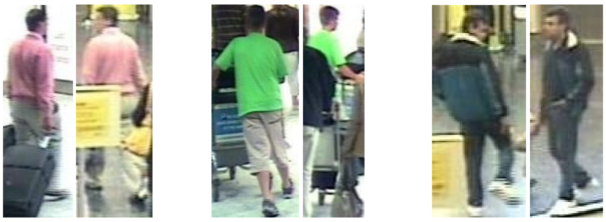


Figure 4.5: Pairs of images showing the same individual taken from the i-LIDS data set. Notice pose and illumination variations, and partial occlusions.

and strong occlusions (see Fig. 4.5). The same experimental set-up as in [139] is used: one image for each person was randomly selected to build the template gallery, while the other images formed the probe gallery. Therefore, the template gallery is composed of 119 images, while the probe gallery has 357 images. The whole procedure is repeated ten times. The folds originally used in [139] are not available, therefore they have been generated randomly.

The re-identification accuracy has been evaluated using the CMC curve defined in Sect. 4.1. The re-identification time has been evaluated using the model described in the same Section.

4.3.2 Results

First, the raw processing time, memory requirements, and re-identification performance of MCMimpl and of its MCD-based version (which in the following is referred to as MCMimpl$^{\text{Dis}}$) are assessed. Then, the trade-off between accuracy and computational time on the real-time application scenario depicted in Sect. 4.1 is evaluated by the model proposed in the same Section. Finally, two critical aspects of the proposed approach are assessed, namely, the number of prototypes and the gallery of individuals used to construct them.

Computational requirements and re-identification accuracy

Processing time and memory requirements of MCMimpl$^{\text{Dis}}$, have been evaluated on a 2.4 GHz CPU, using C# code without any particular optimisation or parallelisation. Results are shown in Table 4.1. Processing times are averaged over ten runs of the experiments, except for VIPeR-632. The average time for prototype construction is reported for four different sizes of the template gallery, corresponding to the four different datasets considered (i-LIDS, VIPeR-316, VIPeR-474, and VIPeR-632). The average total time required for a single run of the experiments is also reported for each dataset: it comprises creation and matching of template and probe descriptors, and also prototypes creation and dissimilarity vectors construction, for MCMimpl$^{\text{Dis}}$.

	MCMimpl	MCMimpl$^{\mathrm{Dis}}$
Avg time for template descriptor creation	93.7 ms$^{(1)}$	17.5 ms
Avg time for probe descriptor creation	6.8 ms$^{(1)}$	17.5 ms
Avg time for prototypes creation, 119 templates	-	2447.3 ms$^{(2)}$
Avg time for prototypes creation, 316 templates	-	6083.2 ms$^{(2)}$
Avg time for prototypes creation, 474 templates	-	12384.8 ms$^{(2)}$
Avg time for prototypes creation, 632 templates	-	16270.7 ms$^{(2)}$
Avg time for dissimilarity vector creation	-	110.3 ms$^{(2)}$
Avg time for a single match	28.6 ms	0.004 ms
Avg total time for a single run (i-LIDS)	2719.1 sec	63.5 sec
Avg total time for a single run (VIPeR-316)	2887.6 sec	87.2 sec
Avg total time for a single run (VIPeR-474)	6521.0 sec	134.5 sec
Avg total time for a single run (VIPeR-632)	11550.6 sec	179.4 sec
Size of the descriptor	96 KB	1.2 KB$^{(2)(3)}$
Size of the prototype gallery	-	48 KB$^{(2)(3)}$

Table 4.1: Comparison of the computational and memory requirements of MCMimpl and MCMimpl$^{\mathrm{Dis}}$. Notes: (1) in MCMimpl, the construction of a template descriptor includes the generation of simulated patches, and thus requires a higher time than the construction of a probe descriptor; (2) these values refer to 150 prototypes for both the considered body parts (torso and legs); (3) 32 bit floating point values.

As expected, MCMimpl$^{\mathrm{Dis}}$ clearly outperforms MCMimpl in terms of processing time and memory usage. In particular, a speed-up of four orders of magnitude is attained for descriptors matching. The average overall time required to perform a run of the experiments is much lower as well, and the difference increases as the size of the template gallery grows.

Regarding re-identification accuracy, the performance in terms of average CMC curve of MCMimpl and MCMimpl$^{\mathrm{Dis}}$ on the four datasets are reported in Fig. 4.6. MCMimpl$^{\mathrm{Dis}}$ attained a worse recognition performance than MCMimpl on i-LIDS and VIPeR-316, that correspond to the smallest template galleries, respectively 117 and 316 templates. However, the accuracy gap diminished on VIPeR-474, that exhibits a larger template gallery, and almost vanishes on VIPeR-632, that corresponds to the largest template gallery, and is this in the most challenging and most realistic scenario. This suggests that, when the number of templates is very high, as in many practical applications, the dissimilarity-based version of a re-identification method obtained through MCD can attain the same performance as the original, not dissimilarity-based method, while requiring much lower computational and storage resources.

Trade-off between accuracy and processing time

The above results show that the dissimilarity-based version of a re-identification method can perform worse than the original one. Here the model of Sect. 4.1 is used to evaluate whether the resulting trade-off between accuracy and processing time can be nevertheless advantageous, in the real-time application scenario described in the same Section. To this aim, the overall re-identification time t_{r} of MCMimpl and MCMimpl$^{\mathrm{Dis}}$ have been evaluated, through the use of Eq. (4.4). The expected rank of Eq. (4.2) is computed from the CMC curves

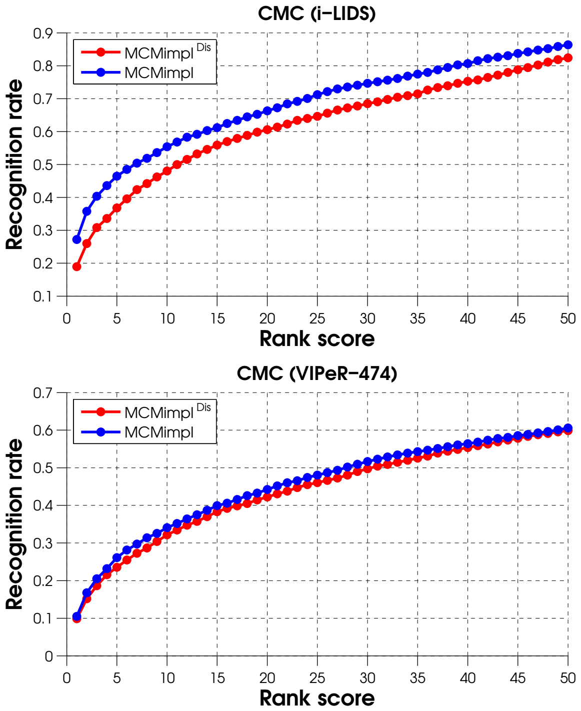


Figure 4.6: CMC curves attained by MCMimpl and MCMimpl$^{\text{Dis}}$ on the four datasets used in the experiments. Note that in the first three plots (from top to bottom, from left to right) average CMC curves over ten runs of the experiments are reported, while the last plot refers to a single run. Note also that the range of rank scores (X axis) is $[1, 50]$ in plots first two plots, and $[1, 100]$ in the last row, since the latter plots correspond to datasets with a larger number of templates.

of Fig. 4.6. To evaluate the time t_{d} required by MCMimpl$^{\text{Dis}}$ for creating one descriptor, both the time needed to build the MCM descriptor, and the time to build the corresponding dissimilarity representation, have been considered.

For the sake of completeness, t_{r} has been evaluated for all the four datasets: the i-LIDS and VIPeR-316 datasets, where the MCMimpl$^{\text{Dis}}$ attained a lower accuracy than MCMimpl; and the VIPeR-474 and VIPeR-632 datasets, were the accuracy of the two methods was similar. The results are reported in Table 4.2.

The overall re-identification time is the sum of two quantities, the processing time t_{p} (i.e., the time required by the system to rank templates in respect to a probe) and the search time t_{s} (i.e., the time spent by the operator to find the individual in the ranked list of templates). As expected, the processing time of MCMimpl$^{\text{Dis}}$ is lower than the one of MCMimpl. The search time of Eq. (4.2) is given by t_{c} (i.e., the average time the operator spends in comparing the probe image with one template image) times the expected rank. The latter turned out to be higher for MCMimpl$^{\text{Dis}}$, on i-LIDS and VIPeR-316, due to the lower accuracy. It was slightly higher also for VIPeR-474 and VIPeR-632, although very close to the one of MCMimpl. This means that the overall re-identification time of MCMimpl$^{\text{Dis}}$ will be lower than the one of MCMimpl, for t_{c} lower than a given value t_{c}^*, and higher for $t_{\text{c}} > t_{\text{c}}^*$.

Accordingly, first the value of t_{c}^* has been computed. Table 4.2 shows that the re-identification time of MCMimpl$^{\text{Dis}}$ is lower, if t_{c} is below about 0.8 seconds for i-LIDS, and 1.3 seconds for VIPeR-316. Since it is likely that in a real-time application scenario like the one considered

	MCMimpl	MCMimpl$^{\mathrm{Dis}}$
i-LIDS		
Processing time t_{p}	3.497 sec	0.128 sec
Search time t_{s} (with $t_{\mathrm{c}} = 0.5$ sec)	10.103 sec	12.203 sec
Re-identification time t_{r} (with $t_{\mathrm{c}} = 0.5$ sec)	13.600 sec	12.331 sec
t_{c}^{*}	0.802 sec	
VIPeR-316		
Processing time t_{p}	9.044 sec	0.129 sec
Search time t_{s} (with $t_{\mathrm{c}} = 0.5$ sec)	13.224 sec	16.601 sec
Re-identification time t_{r} (with $t_{\mathrm{c}} = 0.5$ sec)	21.268 sec	16.730 sec
t_{c}^{*}	1.320 sec	
VIPeR-474		
Processing time t_{p}	13.564 sec	0.129 sec
Search time t_{s} (with $t_{\mathrm{c}} = 0.5$ sec)	42.475 sec	43.084 sec
Re-identification time t_{r} (with $t_{\mathrm{c}} = 0.5$ sec)	56.039 sec	43.213 sec
t_{c}^{*}	11.021 sec	
VIPeR-632		
Processing time t_{p}	18.082 sec	0.130 sec
Search time t_{s} (with $t_{\mathrm{c}} = 0.5$ sec)	55.941 sec	57.700 sec
Re-identification time t_{r} (with $t_{\mathrm{c}} = 0.5$ sec)	74.023 sec	57.830 sec
t_{c}^{*}	5.101 sec	

Table 4.2: Comparison of processing time, search time, and overall re-identification time of MCMimpl versus MCMimpl$^{\mathrm{Dis}}$ (see the text for the details).

here t_{c} is lower than these values, these results show that MCMimpl$^{\mathrm{Dis}}$ can be considered advantageous over MCMimpl, despite the lower accuracy. Note finally that in VIPeR-474 and VIPeR-632 t_{c}^{*} is considerably higher.

The re-identification time t_{r} has been also evaluated, for a realistic reference value of $t_{\mathrm{c}} = 0.5$ seconds. It can be seen that t_{r} is always lower for MCMimpl$^{\mathrm{Dis}}$, and the difference with respect to MCMimpl increases as the template gallery size increases.

Finally, it is worth pointing out that the processing time of MCMimpl$^{\mathrm{Dis}}$, namely the delay between the request of the operator and the response of the system, is almost independent on the template gallery size, and exhibits the very low value of about 0.13 seconds. In contrast, MCMimpl requires a much higher processing time, which grows with the number of templates. This difference is due to the extremely fast matching attained by MCMimpl$^{\mathrm{Dis}}$. Indeed, such high matching speed can be attained by any dissimilarity-based re-identification method based on MCD, as the comparison of dissimilarity vectors is always a fast operation.

To sum up, the above results provide evidence that a dissimilarity-based version of an appearance-based re-identification method can attain an advantageous trade-off between accuracy and processing time.

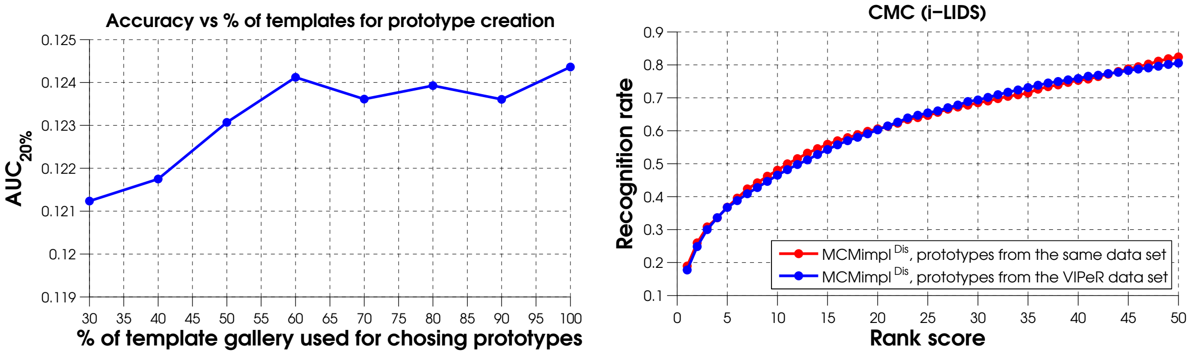


Figure 4.7: (left) Recognition performance of MCMimpl$^{\text{Dis}}$ on the VIPeR-316 dataset, measured as the $AUC_{20\%}$, versus the percentage of the template gallery used to build prototypes. (right) Comparison between the CMC curves of MCMimpl$^{\text{Dis}}$ on the i-ILIDS dataset, attained by constructing the prototypes using either the same dataset, or the VIPeR dataset.

Effect of changing the source and the number of the prototypes

The processing time of MCMimpl$^{\text{Dis}}$, as well as of any dissimilarity-based method obtained via MCD, is affected by prototype construction. This can be a problem, especially in applications where new templates can be added on-line during system operation. For instance, they can correspond to new individuals that are observed by a camera network.

It is thus very interesting to investigate whether the prototype gallery can be constructed using only a subset of the whole template gallery, or even using gallery of individuals *different* than the template gallery. This can avoid to re-build the prototype gallery (and thus, the dissimilarity representation of the existing templates) each time a new template is added to the system. In particular, in the latter case prototypes can be generated off-line, prior to system operation. To this aim, it would be desirable to use a dataset with a wide range of different clothing characteristics.

To assess the performance that can be attained when the prototype gallery is built either from a subset of the template gallery, or from a different gallery, two further experiments have been conducted: 1) an evaluation of the recognition performance in the VIPeR-316 dataset with respect to the percentage of templates used to build the prototype gallery, and 2) the same experiment on i-LIDS of (Fig. 4.6), using this time the VIPeR data set to construct the prototypes, taking into account that VIPeR exhibits a relatively wide range of clothing characteristics.

Results are reported in Fig. 4.7. They show that re-identification accuracy remains almost the same, 1) if at least 60% of the templates in the original gallery are used to construct prototypes (see Fig. 4.7(left)), and, most importantly, 2) if prototypes are constructed using a gallery of individuals different from the template gallery (see Fig. 4.7(right)).

Finally, the accuracy and processing time on the the VIPeR-316 dataset have been evaluated as a function of the number of prototypes per part p. The accuracy has been concisely evaluated as the portion of the area under the CMC curve corresponding to the first 20% of the ranks, denoted as $AUC_{20\%}$. Note that this is the part of the curve of most interest, because it corresponds to the first ranks. The results (shown in Fig. 4.8) provide evidence that the number of prototypes affects performance only slightly.

Fig. 4.8(right) shows that, as the number of prototypes increases, the $AUC_{20\%}$ initially

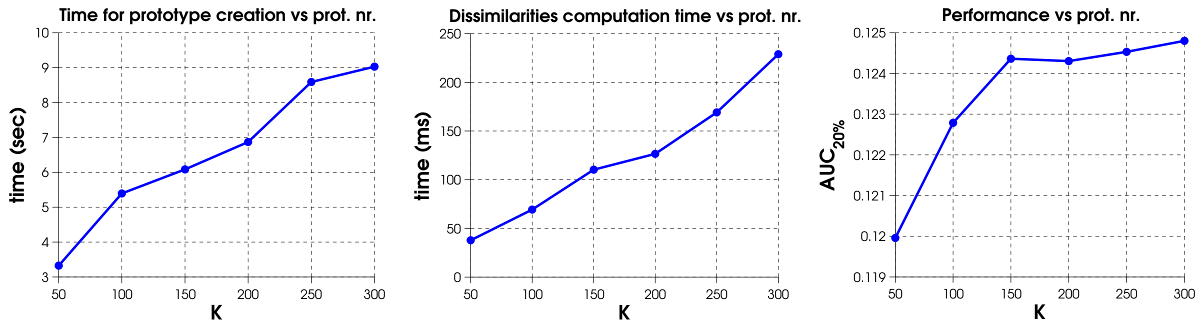


Figure 4.8: (left) Average time for creating prototypes from a dataset of 316 images, versus the number of prototypes per part, K. (center) Average time for computing dissimilarity vectors for a single individual, versus the number of prototypes per part, K. (right) Recognition performance of MCMimpl$^{\text{Dis}}$ on the VIPeR-316 dataset, measured as the $AUC_{20\%}$, versus the number of prototypes per part, K.

grows, then reaches a nearly stable value. This behaviour can be easily explained: once the number of prototypes is enough so that the great part of the distinctive visual characteristics have been captured by different clusters, increasing the number of prototypes has mainly the effect of splitting some of the previous clusters into two or more similar ones. Consequently, no further information is embedded in the new prototypes. On the other hand, Figs. 4.8(left) and 4.8(center) show that increasing the number of prototypes slows down both prototype construction and dissimilarity vector computation. Note that all the plots of Fig. 4.6 correspond to $p = 150$.

4.4 Conclusions

In this Chapter MCD has been used to address the open issue of the computational complexity of person re-identification methods, which has been overlooked so far in the literature. Results have showed that MCD drastically reduces the processing time as well as memory requirements. Also, it can attain a similar accuracy as the original method, especially when the size of the template gallery is high.

Moreover, even if its accuracy is lower, the trade-off attained between accuracy and processing time can be advantageous in terms of the overall re-identification time, in real-time application scenarios. Finally, it has been shown that the visual prototypes needed by a dissimilarity-based method can be constructed either using a subset of the template gallery, or even a different gallery, without affecting re-identification accuracy. This is very relevant for real-time applications as well.

Chapter 5

A state-of-the-art re-identification method based on MCD

In Chapter 4 MCD has been used to speed up an existing re-identification method. In this Chapter, a novel re-identification method based on MCD is presented, which is able to attain state-of-the-art performance with a low computational request. The method, which is described and experimentally evaluated in the following Sections, exploits the representation independence of MCD's prototypes (which are logically and semantically at an higher level than the actual features extracted from objects, as stated in Chapter 3) to combine various kinds of features (both local and global) that look at different aspects of the appearance (e.g., colour, texture). Each feature is responsible for a different set of prototypes, and dissimilarities corresponding to each set are finally concatenated to form the global dissimilarity vector. During matching phase (i.e. when comparing two descriptors), the weighted Euclidean distance of of Eqs. (3.9)-(3.10) takes care of assigning a higher weight to more relevant prototypes, regardless of the underlying features used for each prototype.

The proposed method is able to deal with templates and probes made up of multiple frames, as required in practical scenarios, where typically an entire *track* (i.e., a sequence of frames containing a person) is acquired and processed for each individual seen by the sensor network. Using multiple frames to construct templates should lead to a better recognition performance, as more poses are acquired and a partial occlusion which may happen in a frame may be not present in subsequent frames.

To enable matching of multiple frames (i.e., matching between two sequences of frames, one template sequence and one probe sequence), each frame of the template track is at first matched against each frame of the probe track. The final matching score is then evaluated as a combination of the matching scores between the single template-probe frame pairs. A weighted sum of all the scores is proposed to this aim, where the pairwise matching scores corresponding to similar poses (i.e. when the two frames matched are likely to show a person in the same pose) receive a higher weight.

The rest of the Chapter is organised as follows. Sect. 5.1 illustrates the body model and features used in the descriptor. Details on the application of MCD on this descriptor are then given in Sect. 5.2. The procedure to matching sequences of frames is explained in Sect. 5.3. Then a comprehensive experimental evaluation on two popular benchmark data sets is given in Sect. 5.4. Finally, conclusions are summed up in Sect. 5.5.

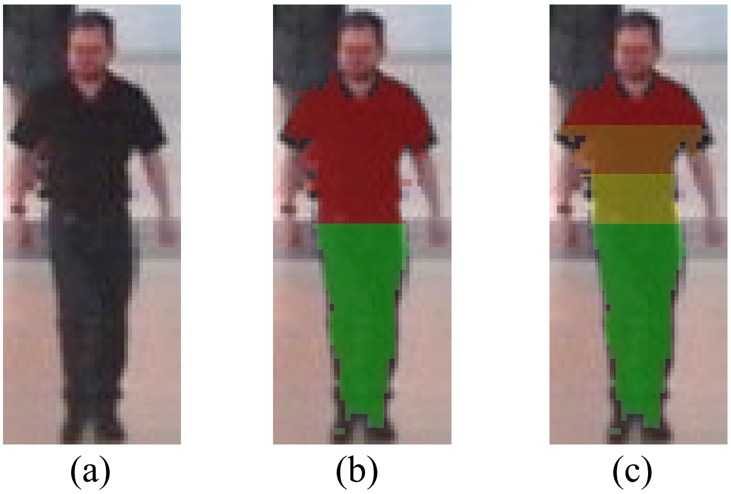


Figure 5.1: Enhanced part subdivision used in the proposed descriptor: (a) the image of an individual; (b) the torso-legs symmetry driven subdivision used in [45]; (c) in the enhanced body subdivision the torso is further subdivided into three body parts of equal height.

5.1 Body model and features used in the descriptor

First, the silhouette of the body is extracted with the same STEL generative model used in Chapter 4.

The proposed method then uses a simple body model which inherits from the symmetry-driven one of [45] (it has been described in Sect. 2.1.2). The original model subdivides the body into two body parts: torso and legs. In the new body model, the torso is further subdivided into three body parts (upper, middle and lower torso) of equal height (see Fig. 5.1). This simple modification of the model of [45] allows to roughly capture the presence of short sleeves, which will result on the presence of skin-like colours in the middle and/or lower torso. It is worth pointing out that, in contrast with more complex body models, such as the one based on pictorial structures used in [28] (see Sect. 2.1.3), the proposed part subdivision is very fast to compute and may be used in real application scenarios.

Concerning the kind of features, five different ones are used, that generally look over different appearance aspects, although some are partly overlapping in this sense. These features are:

1. **RandPatchesHSV**. 100 rectangular patches are sampled at random, and described with the concatenation of H, S, and V colour histograms (32, 24 and 4 bins respectively; note that the V channel is more down-sampled than the other two, as it is more sensitive than the others to brightness variations). The patch width and height were defined as 30% of the width and height of the corresponding part (upper, middle and lower torso) and as 15% of the width and height of the corresponding part (legs).

2. **RandPatchesLBP**. 50 rectangular patches are sampled at random, and described with three rotation-invariant LBP histograms [108], respectively of the H, S and V channels, concatenated to form a single feature vector. The patch size is the same as in the first feature *RandPatchesHSV*.

3. **FCTH**. The *Fuzzy Colour and Texture Histogram* descriptor was originally proposed

for image retrieval [27]. It comprises colour and texture information in one feature vector, based on fuzzy-linking histograms on HSV color space and on the output of Haar Wavelet transforms.

4. **EdgeHistogram**. It is an histogram of the directions of each edge pixel in the image, and one of the suite of MPEG-7 descriptors [121].

5. **SCD**. The *Scalable Colour Descriptor*, another MPEG-7 descriptor, is a colour histogram encoded by a Haar transform [121]. It uses the HSV colors space uniformly quantized to 255 bins, that are subsequently non-uniformly quantised in 64 bits/histogram for a rough representation of the color distribution.

Note that only the first two are local features, while the remaining three are global features extracted from the whole body part.

5.2 Prototype selection and dissimilarity vector creation

MCD is applied to the appearance descriptor above, by separately choosing a prototype set for each kind of feature and each body part, and then concatenating the resulting dissimilarity vectors. To this aim, for each kind of feature it is basically needed to define 1) the prototype selection scheme, and 2) the one-vs-many distance to compute dissimilarities.

Concerning prototype selection, for the two local features *RandPatchesHSV* and *RandPatchesLBP*, the same two-stage clustering scheme of Sect. 4.2.1 is used (with the Band Width parameter of Mean-shift set to $BW = 0.2$), and prototypes are chosen as the patches nearest to the resulting centroids. For the remaining global features *FCTH*, *EdgeHistogram* and *SCD*, k-means is used.

Concerning dissimilarities computation:

- In the case of the two local features *RandPatchesHSV* and *RandPatchesLBP*, each dissimilarity is evaluated as the k-th minimum distance over all the distances between the prototype and each of the components of the body part, with $k = 10$. The pairwise distance between a component and a prototype is evaluated using the Bhattacharyya distance of Eq. (4.7).

- In the case of the three global features *FCTH*, *EdgeHistogram* and *SCD*, the Cosine distance is used, defined as $d(x, y) = 1 - \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$.

Note that all the distances are bounded in $[0, 1]$. This enables the use of the weighted Euclidean distance in the dissimilarity space (Eqs. (3.9)-(3.10)) for matching dissimilarity descriptors.

5.3 Matching between sequences of frames

In practical applications, each person seen by a camera is usually associated to a *track*, i.e. a sequence of rectangular regions of a frame, containing the person. Thus, a real-world re-identification system is likely to have to deal with template and probe *tracks* rather than single frames. Both the template and the probe gallery are therefore made up of sets of frames.

This scenario is often called MvsM (Multiple shots versus Multiple shots) in the literature [45], while the simple one template frame versus one probe frame scenario is called SvsS. In [45], a third scenario is described, the MvsS, where templates are sets of frames while the probe is one single frame. This can correspond to a scenario in which a *continuous* re-identification is performed, i.e. where the person is continuously matched against the templates in real-time, using the currently seen frame as probe.

In the proposed method, the matching in MvsM and MvsS scenarios is performed as follows. For each comparison between one template track and each probe track, at first every frame of the template track is matched against every frame of the probe track using the weighted Euclidean distance of Eqs. (3.9)-(3.10). Then, to compute the final matching score, one of the following approaches is used:

1. score selection, by ranking the minimum, median, or maximum value of the pairwise matching scores;

2. weighted sum of the pairwise matching scores, whose weights are computed with a novel algorithm that takes into account the pose similarity. This algorithm will be described in the following.

5.3.1 Weighting of multiple matches

The proposed algorithm for weighting multiple matches build on the assumption that matches involving similar poses (i.e., frames showing two individuals in a similar pose, e.g. both in frontal pose) should receive an higher weight than matches involving different poses. To estimate the pose of a person from one image is a non-trivial task. However, in this task we are not really interested on the actual person's pose; rather, we are interested in the *similarity of a pair of poses*. In the following, a possible approach to estimate the degree of similarity of two poses given two images is presented.

Consider a pair of frames f_1 and f_2 showing two individuals. Let these two images be subdivided into an equal number of horizontal strips. If the individuals shown in the images are in a different pose (e.g., one in frontal pose, one in lateral pose), the average Hue [1] of each strip should different. E.g., the average H of the strips that contain the head should change from frontal to lateral pose, because of the different amount of skin colour and hair colour in the two cases.

Let S_f be the vector of average values of the Hue channel in each strip of the frame f:

$$\mathbf{s_f} = \begin{bmatrix} \text{avg}(f_{strip=1}) & \ldots & \text{avg}(f_{strip=S}) \end{bmatrix} \tag{5.1}$$

where S is the number of the strips. Following the intuition above, the pose similarity of f_1 and f_2 can be roughly evaluated by considering the differences of the two corresponding vectors of average strip Hue values $\mathbf{s_1}$ and $\mathbf{s_2}$. This difference is evaluated using the Cosine similarity, which is bounded in $[0,1]$ by definition. The pose similarity is thus defined as:

$$\text{pose similarity}(f_1, f_2) = \frac{\sum_i s_{1,i} \cdot s_{2,i}}{\sqrt{\sum_i s_{1,i}^2} \cdot \sqrt{\sum_i s_{2,i}^2}}. \tag{5.2}$$

[1] Hue is the H channel in the HSV colour space, and encodes the actual colour; the other channels are S which encodes the saturation of the colour, and V, which encodes its intensity.

The similarity above is used to compute the weights of each matching pair. Given a template set of frames $T = \{t_i\}$ to compare with a probe set of frames $P = \{p_j\}$, the final matching score is given by:

$$\text{score} = \sum_{i,j} \frac{\left(\text{pose similarity}(t_i, p_j)\right)^C}{\Sigma} \cdot d^{\mathrm{D}}(t_i, p_j), \tag{5.3}$$

$$\Sigma = \sum_{i,j} \left(\text{pose similarity}(t_i, p_j)\right)^C, \tag{5.4}$$

where C is a parameter that governs how much low and high similarities are differentiated, Σ is a normalisation factor such that the sum of the weights is one, and $d^{\mathrm{D}}(t_i, p_j)$ is the matching distance in the dissimilarity space between t_i and p_j, computed by means of the weighted Euclidean distance of Eqs. (3.9)-(3.10).

5.4 Experimental evaluation

In this Section, the proposed method is experimentally evaluated on two common benchmark data sets, which are described in Sect. 5.4.1. All the five features used (see Sect. 5.1) are at first separately tested in Sect. 5.4.2. Then their combination in a single dissimilarity descriptor is evaluated in Sect. 5.4.3. Finally, in Sect. 5.4.4 the score combination rules of Sect. 5.3 are tested, and the performance of the method is compared against current state-of-the-art re-identification methods.

5.4.1 Experimental set-up

The following experimental evaluation has been carried out using two benchmark data sets. One is a set of images taken from the i-LIDS MCTS video dataset (in the following simply denoted as i-LIDS), the other is the CAVIAR4REID data set [28]. Both data sets contain several images per individual, so that the MvsM scenario can be properly simulated.

The i-LIDS data set is the same used in Sect. 4.3. It contains 476 images of 119 different pedestrians, taken at an airport arrival hall from different non-overlapping cameras (see Fig. 4.5). Each pedestrian is associated to at least two images (seven images maximum). With this data set, the MvsM and MvsS scenarios have been tested, using the following procedures for constructing the template and probe galleries:

- Concerning the MvsM scenario, from each individual's N frames in the data set up to $M = 3$ template frames are chosen at random. Up to M frames are then chosen as probe frames for that individual, from the remaining ones. If $N < 2M$, first $M' = \min(N-1, M)$ frames are taken at random as templates, then $M'' = \max(1, N - M')$ frames are taken at random from the remaining ones, as probe frames.

- Concerning the MvsS scenario, from each individual's N frames in the data set up to $M = 3$ template frames are chosen at random. One frame is then chosen as probe frame for that individual, from the remaining ones. If $N < M+1$, first $M' = N-1$ frames are taken at random as template frames, then the remaining frame is taken as probe frame.

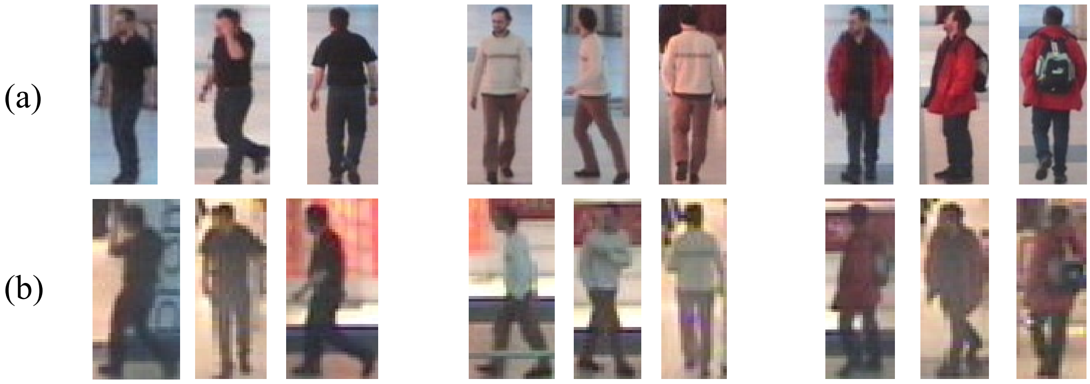


Figure 5.2: Images taken from the CAVIAR4REID data set: (a) images of three individuals from the first camera, and (b) the same individuals seen by the second camera. Notice pose variations, and the strong difference in quality between the images taken by the first camera, which are generally fair, and those taken by the second camera, which are typically blurry and darker.

In both scenarios, prototypes are chosen from the template gallery (Sect. 5.2) and dissimilarity descriptors of the templates and of the probes are computed. Finally, for each probe frame, the template sets of frames are ranked with respect to their similarity to the probe. The procedure is repeated ten times, and the performance is assessed by means of the Cumulative Matching Characteristics curve.

The CAVIAR4REID data set contains tracks of 72 pedestrians taken from two cameras A and B in an indoor shopping centre in Lisbon. To properly simulate a multiple camera scenario, here the 50 pedestrians whose tracks are captured by both cameras are used. Each track is made up of 10 frames. The first camera is used as source for templates, while the second camera is used as source for probes. Despite the relatively small number of individuals, this is a very difficult data set (see Fig. 5.2), mostly due to poor illumination conditions (especially for the second camera) and pose variations. It is thus one of the most realistic data sets currently available. With this data set, only the MvsM scenario was tested, as follows. For each individual, $M = 5$ template frames are chosen at random from camera A, then M probe frames are chosen at random from camera B. Prototypes are chosen from the template gallery (Sect. 5.2) and dissimilarity descriptors of the templates and of the probes are computed. Then, for each probe set of frames, the template sets of frames are ranked with respect to their similarity to the probe, using the method of Sect. 5.3. The procedure is repeated ten times, and the performance is assessed by means of the Cumulative Matching Characteristics curve (Sect. 4.1).

Finally, concerning the parameters of the method, the number of prototypes per part has been set to 200 for *RandPatchesHSV* and *RandPatchesLBP*, and to 100 for *FCTH*, *Edge-Histogram* and *SCD* (in total, 2800 prototypes).

5.4.2 Performance of the single feature sets

The performance attained by using the single feature sets in the three scenarios considered (i-LIDS MvsM, i-LIDS MvsS, CAVIAR4REID MvsM) is reported in Figs. 5.3-5.5. In all the experiments, the final matching score between one template track and one probe track is com-

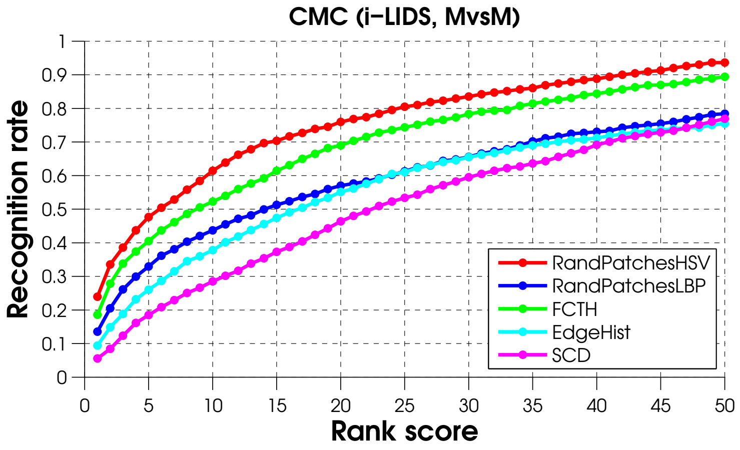


Figure 5.3: CMC curves attained by the single features on the i-LIDS data set (MvsM scenario with M=3).

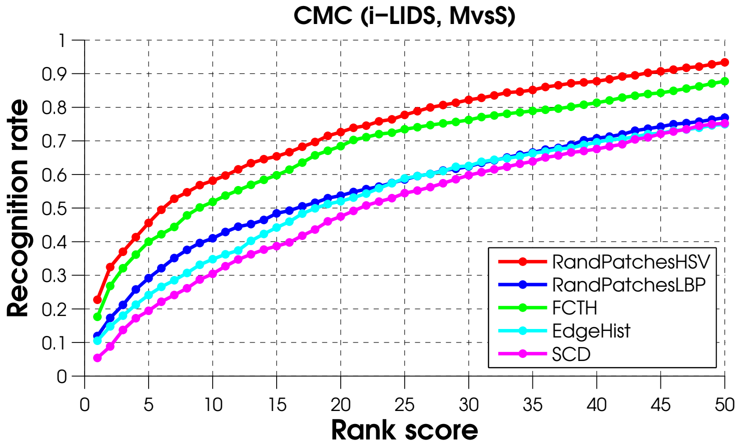


Figure 5.4: CMC curves attained by the single features on the i-LIDS data set (MvsS scenario with M=3).

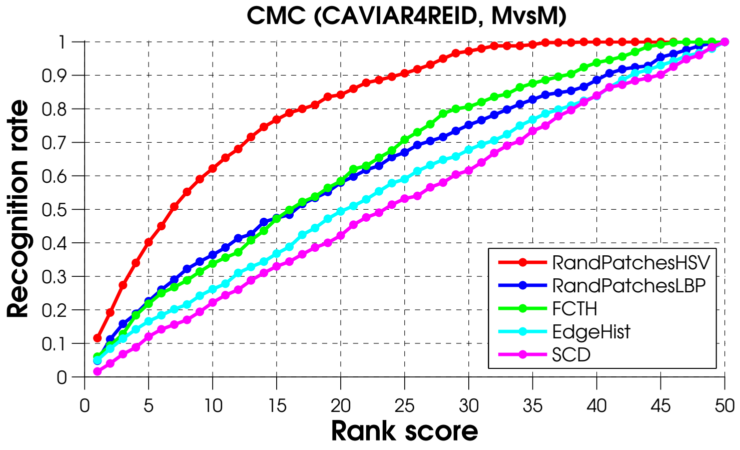


Figure 5.5: CMC curves attained by the single features on the CAVIAR4REID data set (MvsM scenario with M=5).

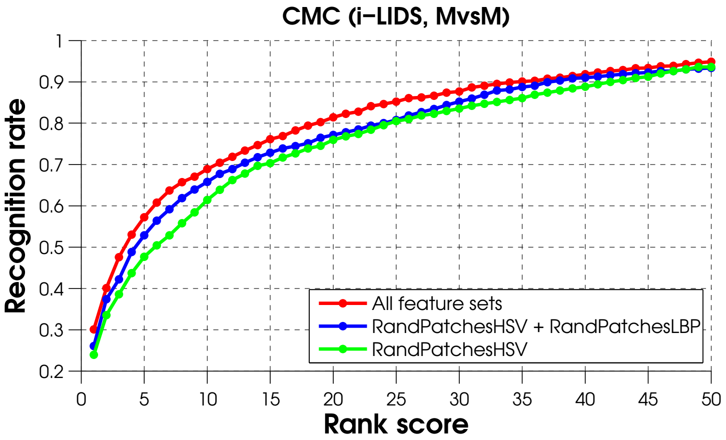


Figure 5.6: CMC curves attained by combining feature sets on the i-LIDS data set (MvsM scenario with M=3). The performance of the best single feature set, *randPatchesHSV*, is also reported for reference.

puted as the median value of the matching scores between each frame of the template track and each frame of the probe track.

Results show that the best performing single feature is *randPatchesHSV*. This was expected, as it looks at colour-related information, and is able to capture local appearance characteristics, which depending on viewpoint and pose may appear in different positions in different images. In contrast, the other colour-related features (*FCTH* and *SCD*) can only capture global colour characteristics. Therefore, they are less robust to viewpoint and pose changes, and accordingly perform worse than *randPatchesHSV*.

Texture-only features (*randPatchesLBP* and *EdgeHist*) perform poorly: indeed, information on textures is likely to be not distinctive enough when used alone. Instead, the performance of re-identification systems that use a combination of colour-related and texture-related local and global features should take advantage of their complementary information. This intuition is studied in the next Section.

5.4.3 Performance of the combination of feature sets

The single features of Sect. 5.1 have been combined by concatenating their corresponding dissimilarity vectors to form a single vector of size 2800 (200 each relative to *randPatchesHSV* and *randPatchesLBP*, 100 each relative to the remaining three feature sets, for each of the four body parts). The performance of the resulting dissimilarity descriptor in the two data set considered is shown in Fig. 5.6-5.8, which also report the performance of the combination of local features only (*randPatchesHSV* and *randPatchesLBP*), and the performance of the best single feature set (*randPatchesHSV*) as reference. In all the methods above, the matching score has been computed as the median value of the scores between each frame of the template track and each frame of the probe track.

Results confirm that combining feature sets that look at different appearance aspects can help on attaining a better performance. This is particularly evident in the i-LIDS data set (Figs. 5.6-5.7), while in the CAVIAR4REID data set the improvement that can be attained by combining features is only modest (Fig. 5.8). Interestingly, while the best performance on the i-LIDS data set is exhibited by the combination of all the five feature sets considered, in the

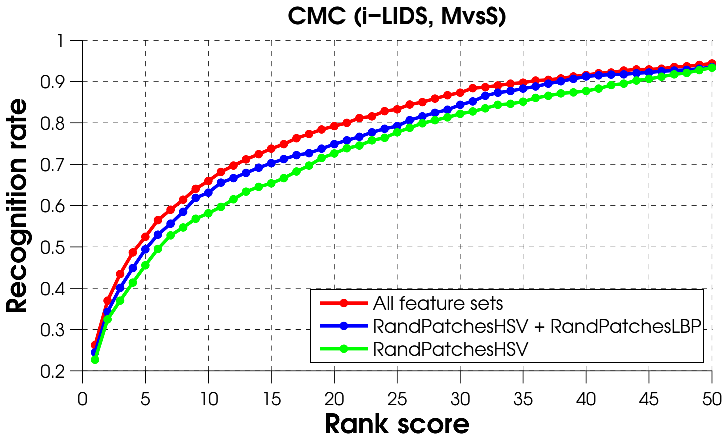


Figure 5.7: CMC curves attained by combining feature sets on the i-LIDS data set (MvsS scenario with M=3). The performance of the best single feature set, *randPatchesHSV*, is also reported for reference.

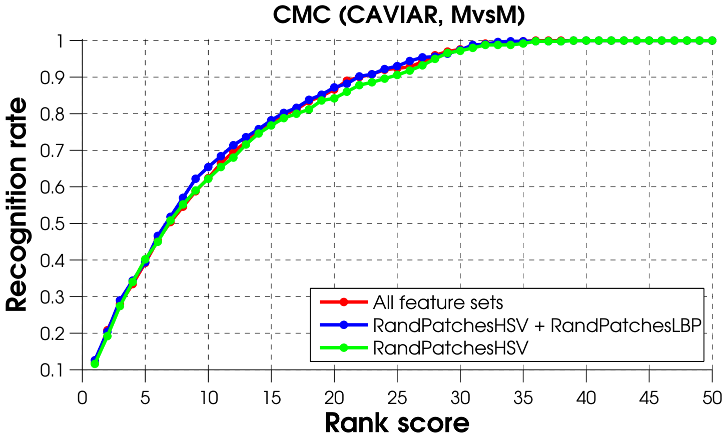


Figure 5.8: CMC curves attained by by combining feature sets on the CAVIAR4REID data set (MvsM scenario with M=5). The performance of the best single feature set, *randPatchesHSV*, is also reported for reference.

CAVIAR4REID data set the best performing combination is the one involving only the two local features *randPatchesHSV* and *randPatchesLBP*. This may be due to the wider variety of poses shown in the images of the CAVIAR4REID, as local features are generally more robust than global features to pose variations.

5.4.4 Performance of score combination rules and comparison with the state of the art

The matching score combination rules of Sect. 5.3 are compared in the following. Specifically, six combination rules have been tested: maximum score, minimum score, median score, and the weighted score of Sect. 5.3.1 with three different values of the parameter C (1, 4 and 16). The attained results are shown in Figs. 5.9-5.11, and include the performance of the median score rule that has been used in the previous Sections (which, as becomes evident from the plots, is not the best combination rule).

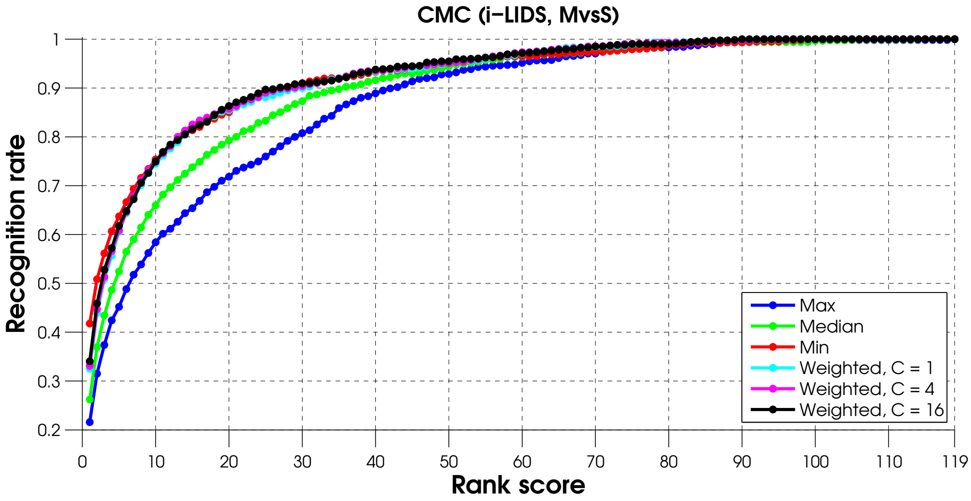


Figure 5.9: CMC curves attained by the best combination of feature sets on the i-LIDS data set (MvsS scenario with M=3) using different matching score combination rules.

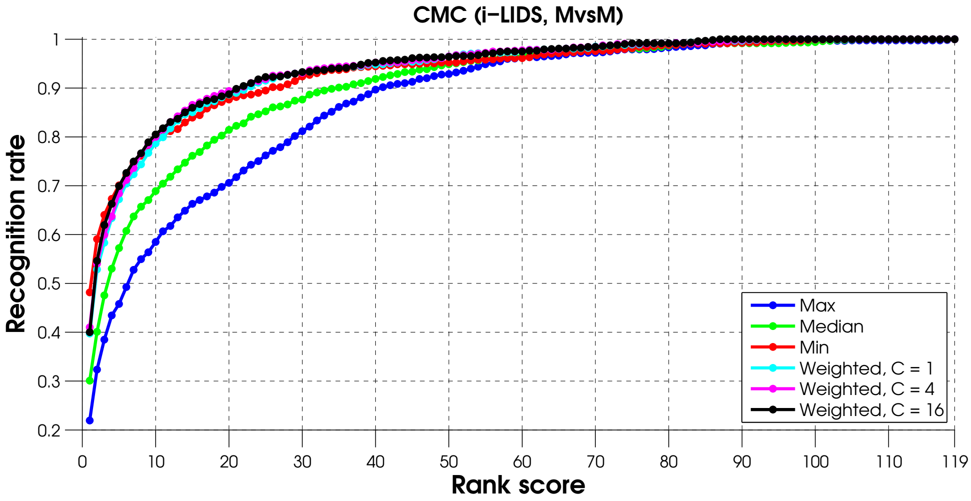


Figure 5.10: CMC curves attained by the best combination of feature sets on the i-LIDS data set (MvsM scenario with M=3) using different matching score combination rules.

Results on the i-LIDS data set are reported in Figs. 5.9-5.10 respectively for the MvsS and MvsM scenario. Note that the range of ranks has been extended to $[1, 119]$ in these plots (being 119 the number of templates), to visualise all the differences of the CMC curves. Fig. 5.11 report the results on the CAVIAR4REID data set.

On the i-LIDS data set (both scenarios), when considering the overall behaviour of the CMC, the best performing score combination rule is the weighted one with $C = 16$, although the difference with respect to the minimum score is modest. This was expected, as the i-LIDS data set shows only slight pose variations (most images are taken from the back), and thus the weighting based on pose similarity does not help much. Note, also, that the minimum score rule performs slightly better than all the other rules when looking at the first ranks of the CMC only.

The weighted rule performs better than the other ones also on the CAVIAR4REID data set, where the best CMC is attained using $C = 1$. In this case, the influence of the C param-

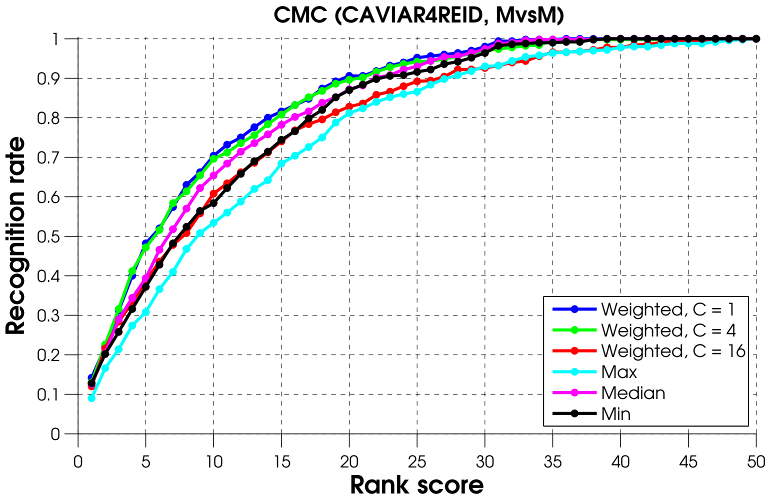


Figure 5.11: CMC curves attained by the best combination of feature sets on the CAVIAR4REID data set (MvsM scenario with M=5) using different matching score combination rules.

eter is higher than in the i-LIDS. Again, this behaviour can be explained by considering the peculiarities of the data sets. In fact, the CAVIAR4REID shows a wide variety of poses for each individual, which explains why weighting the matches with respect to the pose similarity allows for a better performance. Concerning the behaviour with respect to the *C* parameter, it is first important to point out that *C* controls how much the weights differ one another. The more one is confident that the similarity of two poses can be evaluated reliably, the bigger should be the value of *C*. Since many images of the CAVIAR4REID are of very low quality, the reliability of estimation of pose similarity is low, therefore low values of *C* are preferable. This fact is confirmed by Fig. 5.11, that shows a performance that decreases as *C* increases.

Finally, it is worth to note that in all the data sets and scenarios considered the worst performing score combination rule is the maximum score, whose corresponding CMC curve is considerably lower than those attained using the other rules. Among the “classical” score combination rules, in the i-LIDS data set (both MvsS and MvsM scenarios) the minimum score rule, which is the most conservative of the three, performs best and is comparable to the weighted ones. In the CAVIAR4REID data set, instead, the median score rule is the best of the three, although it performs considerably worse than the weighted rules.

The best score combination rule has been compared with two state-of-the-art methods on the two data sets considered. The the Author’s best knowledge, the best performing method, in both data set and in the MvsM scenario, proposed so far in the literature is the one by Cheng et al.[28], denoted as CPS in the following. CPS is basically an enhanced version of another method, SDALF [45]. In place of the symmetry-driven subdivision into torso and legs used by SDALF, it uses a custom Pictorial Structure to subdivide the body in six body parts: chest, head, thighs and legs. This articulated body model has been described in Sect. 2.1.3. From each body part, two kinds of feature are extracted: Maximally Stable Colour Regions (MSCR), which are non-regular regions of homogeneous colours, and a weighted HSV histogram. Details on these features have been given in Sect. 2.2. The second method used for comparison is SDALF [45], which is probably the most widely cited method for person re-identification, and shows the best reported performance for the i-LiDS MvsS scenario.

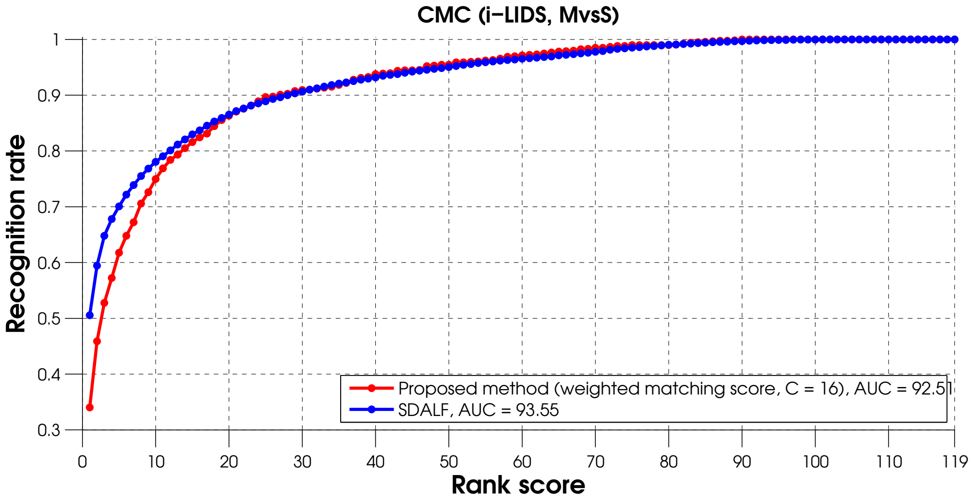


Figure 5.12: Comparison of the best performing score combination rule with SDALF [45], in terms of CMC curve attained on the i-LIDS data set (MvsS scenario).

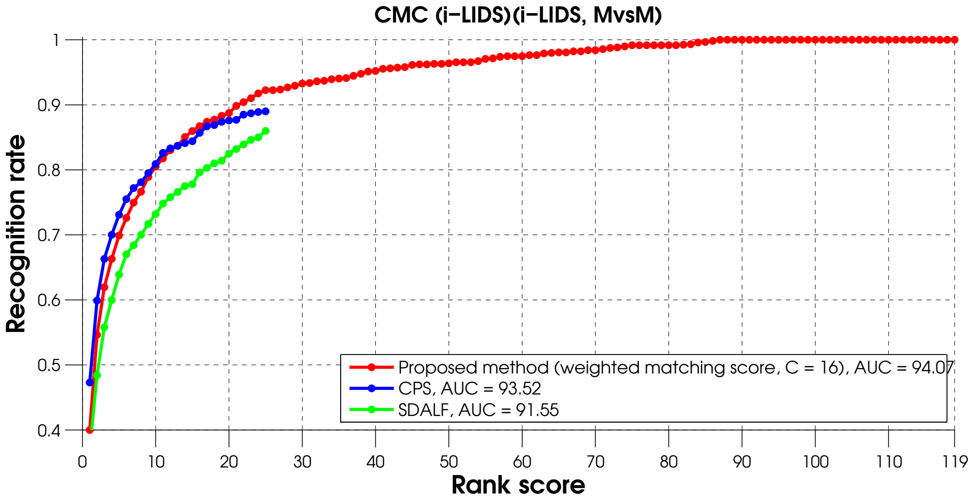


Figure 5.13: Comparison of the best performing score combination rule with two state-of-the art methods, CPS [28] and SDALF [45], in terms of CMC curve attained on the i-LIDS data set (MvsM scenario). Note that the authors of [28] and [45] reported only the first 25 ranks of the CMC, therefore the two plots are truncated at the 25th rank. The area under CMC curve (AUC) is also reported (for CPS and SDALF, the values have been taken from [28], and refer to the full-range CMC).

Figs.5.12 - 5.13 shows the comparisons on the i-LIDS data set respectively for the MvsS and MvsM scenarios. Fig. 5.14 shows the comparison on the CAVIAR4REID data set.

The authors of [28] made only the first 25 ranks of the CMC available for the i-LIDS, therefore the corresponding CMC are truncated. For the same reason the CMC attained by CPS and SDALF on the CAVIAR4REID data sets are truncated at the 30th rank. A useful scalar value to compare CMC curves is the area under the curve (AUC), which has been evaluated on [28] for both CPS and SDALF in both data sets. Therefore, the AUC is also reported in the plots to complete the comparison. Note that all the AUCs are computed using the full-range CMC.

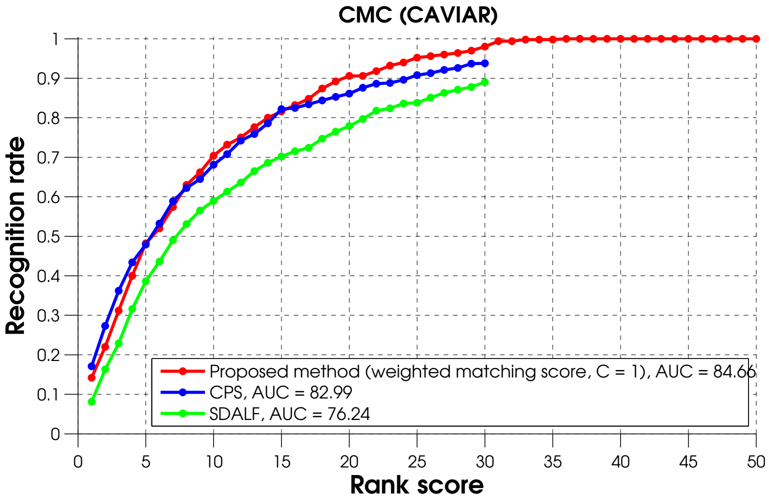


Figure 5.14: Comparison of the best performing score combination rule with two state-of-the art methods, CPS [28] and SDALF [45], in terms of CMC curve attained on the i-LIDS data set. The CMC reported for each method refers to the scenario (MvsM or MvsS) where the method performs better. Note that the authors of [28] and [45] made only the first 30 ranks of the CMC available, therefore the corresponding CMC is truncated at the 30th rank. The area under CMC curve (AUC) is also reported (for CPS and SDALF, the values have been taken from [28], and refer to the full-range CMC).

Considering first the MvsM scenario (Figs. 5.13 - 5.14), in the two data sets considered the proposed method performs worse than CPS considering the cumulative recognition rate of the first ten (i-LIDS) and five (CAVIAR4REID) ranks. However, it shows an overall better performance, and outperforms both CPS and SDALF in terms of AUC. Considering the MvsS scenario (i-LIDS only), the authors of [28] did not report the performance of CPS, therefore the proposed method is compared in Fig. 5.12 against SDALF only. Surprisingly, by confronting Fig. 5.12 with Fig. 5.13 it can be noted that SDALF performs better when only one frame is available per probe individual (MvsS scenario) than in the case of multiple frames per probe available (MvsM). A possible explanation is that the procedure for accumulating frames into a single descriptor used by SDALF does not retain the useful information that can be provided by additional frames. In the MvsS scenario SDALF performs also better than the proposed method, both considering the first ranks only and considering the overall CMC curve and AUC.

The results attained by the proposed method are worth noting especially if one looks at the computational requirements. In fact, the custom pictorial structure used by CPS takes between 15 and 30 seconds *per frame* to be estimated on an i5 2.4GHz CPU (with the C++ code made available by Andriluka et al. [4]), which severely limits any practical usage of CPS. SDALF is faster in descriptor computation: its implementation made available by the authors of SDALF, written partly in C++ and partly in MATLAB, requires about 13 seconds to build one descriptor on the same CPU [116]. Although it can be expected that a pure C++ implementation would be faster, still SDALF descriptor computation appears too slow for a practical implementation on a real-world re-identification system. In contrast, the time required by the proposed descriptor to compute all the feature sets from a single frame (including dissimilarities computation) is about 250 ms, which makes it possible to extract descriptors at 4 fps (frames per second), a value that is enough for most applications. Concerning matching time of one template frame with one probe frame, the authors of CPS [28] do not report any information. However, it should be similar to that of SDALF, which is about 60 ms on the same CPU above and the code made available by the authors. Instead, the matching time of the proposed method (using all the feature sets), is extremely low thanks to the compactness of MCD dissimilarity-based descriptors: about 0.08 ms for a single match. It is worth pointing out that this low matching time enables real-time, on-line re-identification even with big template galleries.

5.5 Conclusions

This Section showed the potential of MCD-based representations for the task of person re-identification. Five different, partially complementary feature sets have been combined together using MCD into a single, compact dissimilarity-based descriptor, which is able to attain state-of-the-art performance on two benchmark data sets and exhibits a very low computational complexity. The method has been applied to an MvsM scenario, i.e., when individuals seen by the camera network are associated to a track instead of a single frame, which is a more realistic setting for person re-identification. In this case, each match between a probe individual with a template individual involves more than two frames. In contrast, the MCD matching procedure proposed in Sect. 3.3 has been designed to match pairs of frames (SvsS scenario). To extend the MCD matching procedure to the MvsM scenario, each frame in the probe track is at first a compared with each frame in the template track, then the resulting scores are combined using a novel technique, which is based on pose similarity estimation.

Experimental results have confirmed the superiority of this technique with respect to other classical score combination rules such as the maximum, minimum or median score. In addition, the proposed method has been compared with two existing techniques, namely CPS [28] and SDALF [45] that have attained the highest performance so far in the data sets considered. With respect to these techniques, the proposed method has a far lower computational complexity (especially with respect to CPS), thanks of the simplicity of the feature sets used, and of the compactness of MCD descriptors. It also performs better, in the MvsM scenario, than SDALF and CPS, when considering the overall CMC, attaining a higher Area Under Curve (AUC) at the expense of a lower recognition performance in the first ranks. On the other hand, in the MvsS scenario the performance attained is lower than that of SDALF. While the motivation of this behaviour has to be investigated, still it is worth pointing out that the MvsS scenario is less realistic than MvsM: in fact, typically more than one frame of the probe individual is available in real-world scenarios, as they are usually extracted from of a video-sequence. Therefore, it is more important to achieve a better performance in the more realistic MvsM scenario.

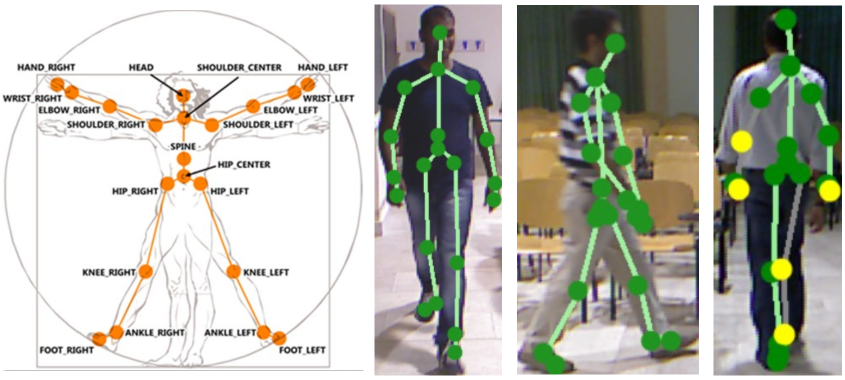
Chapter 6

Using MCD for multi-modal re-identification

As shown in Sect. 2.4, there are various cues alternative to the clothing appearance (e.g., gait, remote face, anthropometry) that, in principle, can be used to address the task of person re-identification. They generally pose strong constraints to the application scenario due to their requirements (e.g., precise silhouette alignment in the case of gait, near frontal pose in the case of remote face). However, the recent introduction of integrated video and range (RGB-D) sensors like the Kinect, and the development of techniques that use the additional range information to estimate the pose and skeleton of individuals (Fig. 2.6), has made it possible to extract various anthropometric measurements, like the height or the average arm length of a person. Although these measurements have a low discriminant capability, their combination with the clothing appearance is likely to be advantageous for a person re-identification system. In this Chapter, the multi-modal approach to person re-identification proposed in Sect. 3.5 is used to develop and experimentally evaluate a multimodal re-identification system, based on the combination of appearance and anthropometrics.

Three anthropometric measures (height, arm length, leg length) are extracted from depth information, from a Kinect device. Similarly to [10], to evaluate them the joint estimation capabilities of the Kinect SDK are used, by measuring distances between selected joints. Differently to [10], only anthropometric measures that could be extracted from the Kinect skeleton with a good degree of reliability, and in an unconstrained pose setting, are used here, to make them suitable to real-world re-identification scenarios. These measures are separately combined with two appearance descriptors: SDALF [45], and a novel descriptor based on MCMimpl (see Chapter 4), using an enhanced body part subdivision.

To the Author's best knowledge, the proposed approach is the first one that exploits multiple modalities for person re-identification, and is among the few ones based on the emerging RGB-D technology. It is worth pointing out that the technology adopted by the Kinect device set two limits on its usability in video-surveillance. The first one is that it can detect objects within a maximum distance of about 5–6 mt (depending on environmental conditions), which is relatively small for the considered task. The second limit is intrinsic to the adopted technology: due to the use of IR projectors and sensors to build depth maps, the Kinect can not be used outdoor, because of the interference in the IR band caused by sun



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

Figure 6.1: (a) The 20 skeletal points tracked by the Kinect SDK in the classical representation of the Vitruvian Man. (b–d) Depending on he degree of confidence of the estimation of the points position, the Kinect SDK distinguishes between *good* (in green) or *inferred* (in yellow) points, the latter being less reliable than the former. Some of the 20 points could also not be tracked at all, depending on the pose: e.g., in (d) only 18 points are actually being tracked, while the right ankle and foot are missing.

light. Both limits can be overcome in real applications, by using more sophisticated sensors, and a non IR-based technology.

After a description of the overall pipeline (Sect. 6.1), SDALF and MCMimpl descriptors are briefly described in Sect. 6.2. In Sect. 6.3 the three considered anthropometric measurements and their extraction are described. The application of MCD framework to the different modalities is described in Sect. 6.4. The approach is experimentally evaluated on a novel data set acquired using Kinect cameras, in Sect. 6.5. Finally, conclusions are drawn in Sect. 6.6.

6.1 Acquisition and re-identification pipeline

The proposed method has been designed to work with video sequences, which is a realistic setting for re-identification systems. The template gallery is thus made up of one *track* per individual (i.e., the sequence of rectangular regions of the frames that contain a given individual). The re-identification task consists of ranking a gallery of template tracks with respect to their similarity to a probe track where an individual of interest appears. The tracking and segmentation capabilities of the Kinect SDK have been exploited to track individuals in video sequences, and separate their silhouette from the background. If an individual is within the above mentioned range of 5–6 mt to the sensor, the Kinect SDK also constructs a *skeleton* for each frame, made up of at most 20 skeletal 3D points, and provides their coordinates (in meters). The actual number of skeletal points depends on the individual's pose and distance to the sensor (see Fig. 6.1).

Each tracked individual I is associated to a sequence of frames $f_{I,i}$, together with the corresponding segmentation masks $m_{I,i}$ and skeleton points. The frames and the skeleton points are used to extract one clothing appearance descriptor for each frame, and the skeleton points are used to estimate anthropometric measures.

6.2 Clothing appearance

One appearance descriptor is extracted for each frame $f_{I,i}$ and corresponding mask $m_{I,i}$, using two different descriptors: SDALF [45] and a modified version of MCMimpl (see Sect. 4.2).

SDALF uses a two body-part subdivision into *torso* and *legs* (more details can be found in Sect. 2.1.2), discarding the image region corresponding to the head (Fig. 6.2-a). The horizontal axes that separate the two body regions are found by looking at symmetry and anti-symmetry properties of silhouette colour and shape. From each body part, *Maximally Stable Colour Regions* (MSCR) and *Recurrent High-Structured Patches* (RHSP) are extracted. MSCR are non-regular regions of homogeneous colours, and are found via agglomerative clustering; each MSCR is represented by its area, centroid, second moment matrix and average colour, resulting in a 9-dimensional vector. RHSP are rectangular patches made up of recurrent, repeated patterns; each one is represented by a rotation-invariant LBP histogram. Both MSCR and RHSP are sampled mainly around the vertical axis of symmetry of each body part. Finally, an HSV histogram is extracted from each body part. The histogram is weighted, so that pixels of the periphery of the body receive less weight than pixels near the vertical axis. The two weighted histograms are then concatenated to form a single feature vector.

In the formulation of Sect. 4.2, MCMimpl randomly extracts multiple, possibly overlapping patches of from two body parts, obtained via the SDALF symmetry-driven subdivision into torso and legs. Each patch is represented via an HSV colour histogram. Furthermore, for each real patch a number of *simulated* ones is generated, by varying brightness and contrast, in order to increase robustness to illumination changes. In this implementation, the original part subdivision of MCMimpl has been changed, exploiting the skeleton points extracted by the Kinect SDK (Fig. 2.6). Four different body parts, *upper torso*, *lower torso*,*upper legs*, and *lower legs* (Fig. 6.2-b)) have been defined as follows.

The region containing the torso is first located as the portion of the image between the y coordinates of shoulder and hip centres. The pixels of the mask corresponding to the first half of such region are considered as the upper torso, and the other ones are considered as the lower torso. The mask pixels between the coordinate of the hip center and the average of the two y coordinates of the knees (or the y coordinate of the visible knee, if only one is detected), define the upper legs region. Finally, the mask pixels between the average y coordinate of the knees and the bottom of the mask define the lower torso region. Note that only points that can be detected from *any* body pose are used to perform body part subdivision (Fig. 2.6-b,c,d). In the following, MCMimpl will denote the enhanced MCMimpl descriptor. In terms of multiple parts/multiple components representation, the SDALF descriptor is made up of $M = 5$ sets of components (see Fig. 6.2-a.2): the first four are made up of the MSCR and RHSP feature vectors extracted from torso and legs, and the fifth one is obtained as the concatenation of the two HSV colour histograms of torso and legs. The MCMimpl descriptor is made up of $M = 4$ sets of components, one for each body part (see Fig- 6.2-b.2).

6.3 Anthropometric measures

For each frame $f_{I,i}$, anthropometric measures are extracted from the corresponding skeleton provided by the Kinect SDK (Fig. 6.1-a). In principle, all the pairwise combinations of such measures can be extracted from such skeletal points. However, depending on individual's pose some skeletal points may be estimated unreliably (see Fig. 6.1-d), or may be not

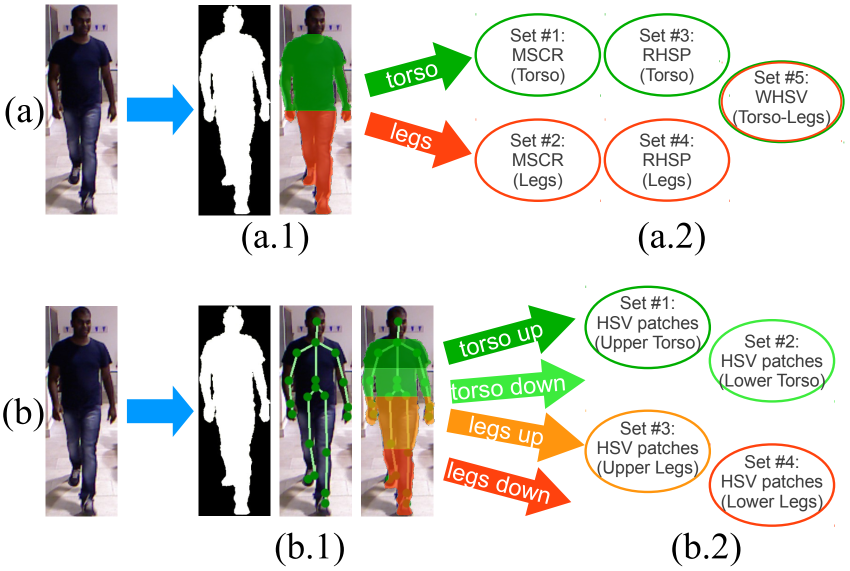

(a)
torso
legs
Set #1: MSCR (Torso)
Set #3: RHSP (Torso)
Set #5: WHSV (Torso-Legs)
Set #2: MSCR (Legs)
Set #4: RHSP (Legs)
(a.1)
(a.2)
(b)
torso up
torso down
legs up
legs down
Set #1: HSV patches (Upper Torso)
Set #2: HSV patches (Lower Torso)
Set #3: HSV patches (Upper Legs)
Set #4: HSV patches (Lower Legs)
(b.1)
(b.2)

Figure 6.2: SDALF (a) and MCMimpl (b) descriptors, both based on a multiple parts/multiple components representation. SDALF subdivides body into torso (disregarding the head) and legs parts (a.1); from each part, RHSP and MSCR local features are extracted, and the weighted HSV histograms of each part are concatenated, leading to five sets of components (a.2). MCMimpl exploits the skeletal points provided by the Kinect SDK to separate body into four parts (b.1); from each body part, a set of partly overlapping patches is randomly extracted, leading to four sets of components (b.2).

extracted at all. Thus, after preliminary experiments, only measures in the vertical direction have been taken into consideration, which are not affected by body pose, and usually correspond to points that can be reliably tracked. Among them, in this work the following three anthropometric measures have been selected, all expressed in meters:

1. individual's *height* a_{height}, evaluated as the distance between the (y,z) coordinates of the top-most and bottom-most points of body silhouette;

2. *average arm length* a_{arm}, evaluated as the sum of the distances between the shoulder and elbow points, between the elbow and wrist points, and between the wrist and hand points, averaged over the two arms (if only one arm is tracked, the measure is taken from that arm);

3. *average leg length* a_{leg}, evaluated as the sum of the distances between the hip and knee points, between the knee and ankle points, and between the ankle and the foot points, averaged on the two legs (if only one leg is tracked, the measure is taken from that leg);

To represent them in an uniform range, a_{height}, a_{arm} and a_{leg} are linearly normalized in the range $[0,1]$:

$$\overline{a}_{\text{height}} = \frac{a_{\text{height}} - \min a_{\text{height}}}{\max a_{\text{height}} - \min a_{\text{height}}} \tag{6.1}$$

$$\overline{a}_{\text{arm}} = \frac{a_{\text{arm}} - \min a_{\text{arm}}}{\max a_{\text{arm}} - \min a_{\text{arm}}} \tag{6.2}$$

$$\overline{a}_{\text{leg}} = \frac{a_{\text{leg}} - \min a_{\text{leg}}}{\max a_{\text{leg}} - \min a_{\text{leg}}} \tag{6.3}$$

where the maximum and minimum values of each measure are obtained from the template gallery. Each anthropometric measure is finally represented as a descriptor corresponding to a single set of components, in which each set is made up of a single scalar component (the normalized value of the corresponding measure):

$$\mathbf{v}_{\text{height}} = [\overline{a}_{\text{height}}], \quad \mathbf{v}_{\text{arm}} = [\overline{a}_{\text{arm}}], \quad \mathbf{v}_{\text{leg}} = [\overline{a}_{\text{leg}}] \tag{6.4}$$

6.4 Application of MCD

The MCD framework has been applied to the above appearance and anthropometric descriptors, to obtain a multi-modal dissimilarity-based representation. The same two-stage clustering scheme of Sect. 4.2.1 has been used to select 200 prototypes of each body part in the case of the appearance descriptors MCMimpl and SDALF, and 8 prototypes in the case of the anthropometric measurements. The influence of varying the number of prototypes will be studied in Sect. 6.5. Dissimilarities are computed by taking the k-th minimum distance over all the distances between the prototype and each of the components of the body part,

$$d(X,p) = k\text{-th}\min_{x\in X} \|x - p\| \tag{6.5}$$

with $k = 10$.

The same distance metric $\|\cdot\|$ between components used for matching SDALF and MCMimpl appearance descriptors is used (the reader is referred to [45] and Sect. 4.2 for further details). For the three anthropometric modalities, each of which is represented as a scalar value in $[0,1]$, $\|\cdot\|$ has been defined as the absolute value of the difference between two measurements.

6.5 Experimental evaluation

To evaluate the proposed multi-modal re-identification system, a data set of video sequences has been acquired, each showing an individual walking in different indoor environments, using Kinect cameras. A novel data set was needed, as to the best of the Author's knowledge no benchmark data sets for re-identification with both RGB and Depth information are currently available. In the following, this data set is referred to as KinectREID. Details on this data set are provided in Sect. 6.5.1.

Two different kinds of experiments have been carried out. First, the re-identification performance of each single modality (appearance-based and anthropometric-based) has been assessed, in Sect. 6.5.2. The original MCMimpl and SDALF descriptors have been evaluated separately, and compared with the corresponding MCD-based methods. Similarly, the performance of each anthropometric measure has been separately assessed. Then, the performance of two versions of the proposed multi-modal re-identification system has been assessed (Sect. 6.5.3), obtained by combining an appearance descriptor (either MCMimpl or SDALF) with anthropometric measures. In this step, the MCD-based fusion approach has been also compared with common score-level fusion rules.

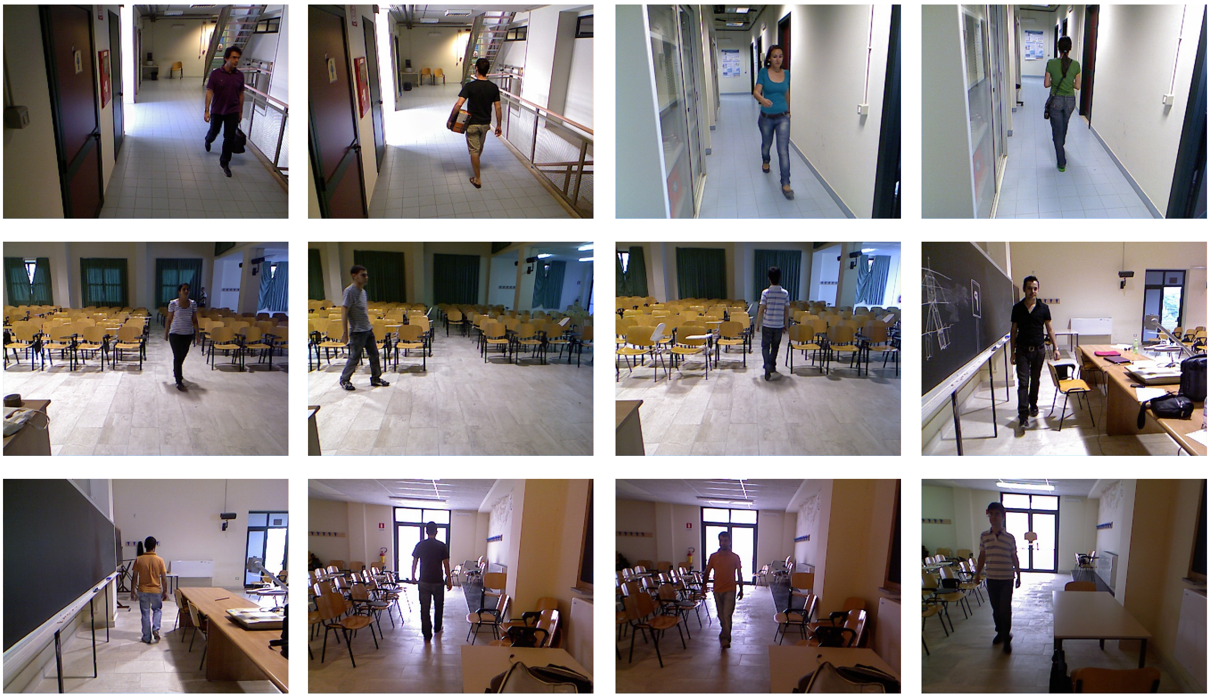


Figure 6.3: Examples of some frames taken from the KinectREID data set. Note the different view points, locations, poses and illumination conditions.

6.5.1 Data set and experimental setup

The KinectREID data set was acquired using different Kinect cameras, and consists of video sequences of 80 different individuals taken at two different locations (the corridors of the department, and a large lecture hall), under different lighting conditions and various view points, including near-frontal views, near-backward views, and lateral views. Individuals were requested to walk normally, and some of them carried bags or other accessories. Each person was associated to 2 to 7 different video sequences. Some examples of frames of the KinectREID data set are shown in Fig. 6.3.

All the experiments have been carried out as follows. First, one video sequence per individual has been randomly chosen as the corresponding template. The remaining video sequences are used as probes. Then, for each template and each probe, 10 frames have been chosen at random among the ones in which the whole body is visible, and the Kinect had been able to extract the skeleton. Since both templates and probes are image sequences, this kind of set-up is often called *multiple shots vs. multiple shots* (MvsM) scenario [45].

Each template has been then ranked with respect to its similarity to each probe. For each comparison, the distance from each pair of template and probe frames has been measured using the distance of Eqs. (3.9)-(3.10). The median value of all the distance has then been taken as final similarity measure. The above procedure has been repeated 20 times, and re-identification performance has been assessed in terms of Cumulative Matching Characteristics (CMC) curve, that is, the average probability of finding the correct match within the first n ranks, with n ranging from 1 to 80.

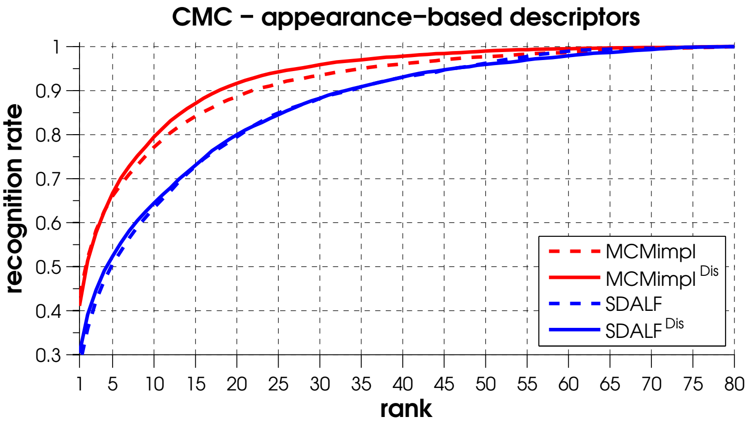


Figure 6.4: Recognition performance, in terms of CMC curve, attained by mono-modal systems on the KinectREID data set by MCMimpl, SDALF, and their corresponding dissimilarity-based versions MCMimpl$^{\mathrm{Dis}}$ and SDALF$^{\mathrm{Dis}}$ (20 runs of the experiments).

6.5.2 Performance using single modalities

The recognition performance attained by mono-modal systems has been first evaluated, both for appearance and anthropometric descriptors.

The performance of the appearance-based descriptors SDALF and MCMimpl, has been assessed using both their original version, and their dissimilarity-based version obtained through MCD (see Sect. 6.4), denoted respectively as SDALF$^{\mathrm{Dis}}$ and MCMimpl$^{\mathrm{Dis}}$. Fig. 6.4 shows the corresponding CMC curves. The recognition performance of the MCD-based methods is comparable or higher to the one of the corresponding original methods, despite the fact that MCD was originally proposed to speed up re-identification methods to make them suitable to real-world scenarios, rather than improving their recognition performance.

As a reference, also the performance of anthropometric-based descriptors has been evaluated. Similarly to appearance descriptors, each MCD-based descriptor has been compared with a baseline method, in which matching distance is computed as the absolute difference between the normalized values of the corresponding measure. Results are shown in Fig. 6.5. As one could expect, anthropometric modalities, when used alone, provide weak discriminant capability. This is reflected by the low performance with respect to appearance modality. Among the three measures, height turned out to be the most discriminant one.

Effect of changing the number of prototypes

The performance of the MCD-based methods may be affected by the choice of the number of prototypes per part K. Note that K also affects computational cost (see Chapter 4): in fact, the more prototypes are used, the longer it will take to compute dissimilarities. To assess the impact of this parameter on the performance, the normalized area under the first 20% ranks of the CMC curve (denoted as AUC$_{20\%}$) with respect to K has been evaluated. Only the top ranks have been considered, since they are the ones of most practical interest. Results attained by SDALF$^{\mathrm{Dis}}$ and MCMimpl$^{\mathrm{Dis}}$ are shown in Fig. 6.6. The AUC$_{20\%}$ attained by the original methods SDALF and MCMimpl is reported with dashed lines. Results attained using anthropometric modalities, compared with the corresponding baseline, are shown in Fig. 6.7. In all the modalities, the performance initially grows as the number of prototypes

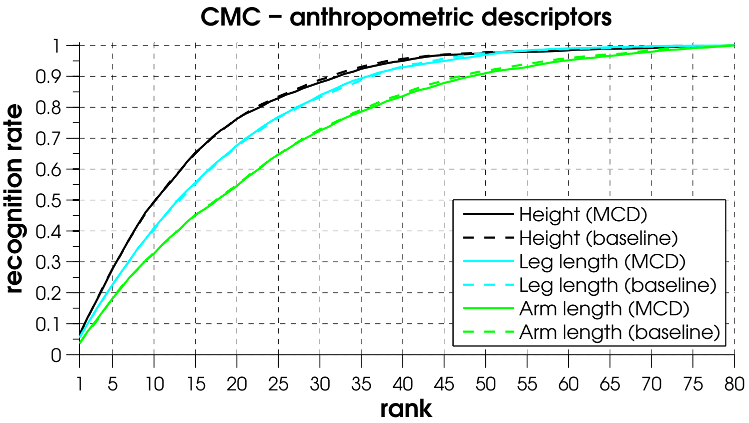


Figure 6.5: Recognition performance, in terms of CMC curve, attained by mono-modal systems on the KinectREID data set by the three considered anthropometric measures, using their baseline (the matching distance is the absolute difference of the normalized values of that measure) and MCD versions (average CMC curve over 20 runs of the experiments).

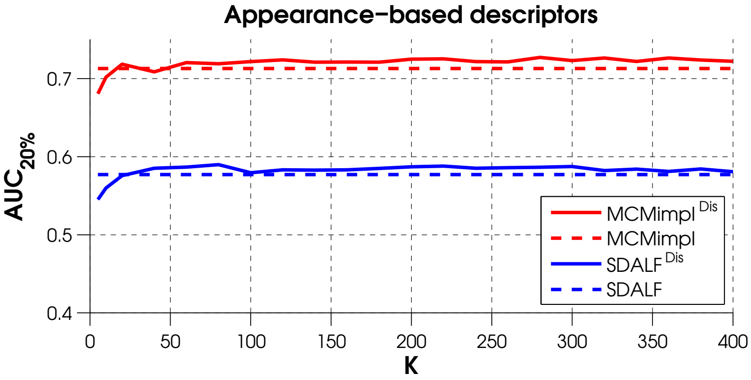


Figure 6.6: Recognition performance, in terms of normalized $\mathrm{AUC}_{20\%}$ (averaged over 20 runs of the experiments), versus the number of prototypes per part K, attained on the KinectREID data set by MCMimpl, SDALF, and their corresponding dissimilarity-based versions $\mathrm{MCMimpl}^{\mathrm{Dis}}$ and $\mathrm{SDALF}^{\mathrm{Dis}}$.

increases, then rapidly reaches a nearly stable value. This makes it relatively easy to select a value of c that gives good performance, or to find a reasonable trade-off between performance and processing time. Note that a similar behaviour has been reported and discussed in more detail in Chapter 4.

6.5.3 Performance of multi-modal systems

This Section assesses the performance that can be obtained by the multi-modal systems of Sect. 6.5.2, using both the MCD-based fusion approach, and the standard score-level fusion approach. The latter has been implemented using several well known fusion rules: minimum, maximum, product, and average of the individual scores. Note that the extended MCD framework can be used to combine only dissimilarity-based descriptors, while score-level fusion can be performed using any descriptor. Nevertheless, only dissimilarity-based

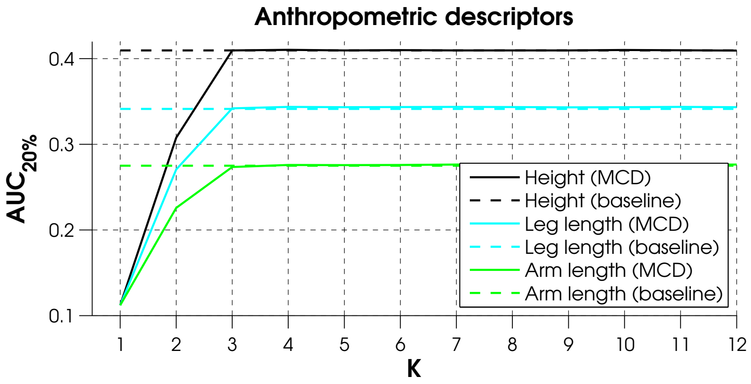


Figure 6.7: Recognition performance, in terms of normalized $AUC_{20\%}$ (averaged over 20 runs of the experiments), versus the number of prototypes per part K, attained on the KinectREID data set by the three considered anthropometric measures, using their baseline (the matching distance is the absolute difference of the normalized values of that measure) and MCD versions.

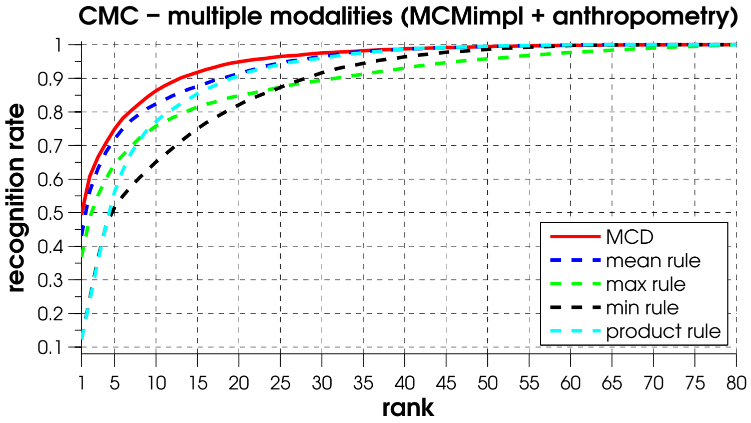


Figure 6.8: Recognition performance, in terms of CMC curve, attained by multi-modal systems, with MCD-based fusion and score-level fusion (maximum, minimum, product, and average rules): $MCMimpl^{Dis}$ and the three anthropometric modalities (height, arm length, leg length).

descriptors are used also in the experiments on score-level fusion, since they outperformed the corresponding non-dissimilarity-based ones.

Figs. 6.8 - 6.9 show respectively the performance attained by combining $MCMimpl^{Dis}$ and $SDALF^{Dis}$ with the three considered anthropometric measures. To better visualise the differences in performance, in the first ten rows of Table 6.1, the cumulative recognition rate, taken from the CMC curves of Figs. 6.8 - 6.9, is also reported for five selected ranks.

As one can see, the combination of different, heterogeneous modalities made it possible to attain higher performance with respect to single modalities. Moreover, the MCD-based fusion approach outperforms the score-level fusion approach (first and sixth row of Table 6.1), except for higher ranks (see Table 6.1, last column). In particular, the first-rank recognition rate of the former is about 6% higher than the latter, when the best fusion rule (average) is used. The best performance overall has been attained when MCMimpl is used as the appear-

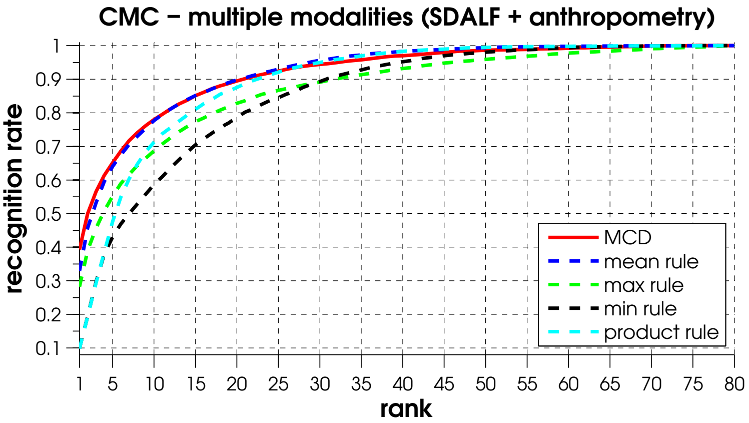


Figure 6.9: (a) Recognition performance, in terms of CMC curve, attained by multi-modal systems, with MCD-based fusion and score-level fusion (maximum, minimum, product, and average rules): SDALF$^{\text{Dis}}$ and the three anthropometric modalities.

	fusion rule	rank 1	rank 5	rank 10	rank 20	rank 40
appearance (MCMimpl) and anthropometry	MCD	**0.495**	**0.749**	**0.863**	**0.949**	**0.987**
	mean rule	0.431	0.720	0.824	0.914	**0.987**
	max rule	0.367	0.645	0.758	0.848	0.937
	min rule	0.122	0.516	0.651	0.821	0.964
	prod. rule	0.112	0.562	0.771	0.9072	0.987
appearance (SDALF) and anthropometry	MCD	0.393	0.652	0.780	0.895	0.972
	mean rule	0.329	0.641	0.777	0.898	0.983
	max rule	0.282	0.552	0.687	0.828	0.932
	min rule	0.099	0.430	0.587	0.787	0.952
	prod. rule	0.098	0.476	0.713	0.876	0.983
appearance only, MCMimpl$^{\text{Dis}}$		0.411	0.667	0.795	0.919	0.978
appearance only, SDALF$^{\text{Dis}}$		0.303	0.524	0.644	0.800	0.931

Table 6.1: Cumulative recognition rates attained by the multi-modal systems, for five selected ranks, taken from the CMC curves of Figs. 6.8 - 6.9. The best recognition rate for each rank is highlighted in bold.

ance descriptor (first row of Table 6.1). These results provide evidence that the MCD-based fusion approach can outperform the standard score-level fusion one, in the specific task of multi-modal person re-identification.

To highlight the gain that can be achieved using multiple modalities, in Fig. 6.10 the performance of the individual appearance-based descriptors is compared to the one attained by their MCD-based fusion with anthropometric measures. The exact values of the cumulative recognition rate attained by individual appearance-based descriptors are also reported in Table 6.1 (last two rows), for five selected ranks. It can be seen that multi-modal fusion

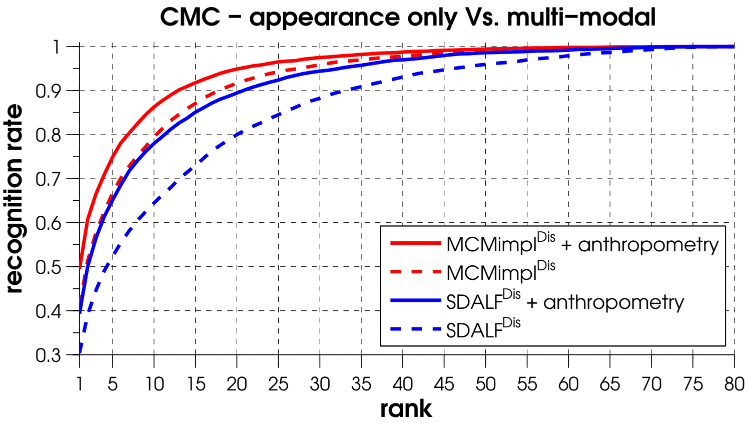


Figure 6.10: Performance, in terms of Cumulative Matching Characteristic curve, of the two MCD-based combinations of appearance and anthropometric modalities, compared with the performance of MCMimpl$^{\mathrm{Dis}}$ and SDALF$^{\mathrm{Dis}}$, on the KinectREID data set.

of appearance and anthropometric descriptors improved the first-rank recognition rate of MCMimpl$^{\mathrm{Dis}}$ and SDALF$^{\mathrm{Dis}}$ by 8% and 9%, respectively. Similarly, the cumulative recognition rate at rank five increases by 9% in the case of MCMimpl$^{\mathrm{Dis}}$, and 12% in the case of SDALF$^{\mathrm{Dis}}$. The improvement becomes lower for higher ranks, that are however of less interest in practical scenarios. Note also that, by definition, the CMC curve attains a 100% recognition rate at the highest rank.

6.6 Conclusions

In this Chapter, a MCD-based re-identification approach that uses a combination of clothing appearance with three different anthropometric traits has been proposed and experimentally evaluated, exploiting the depth information provided by RGB-D sensors. To the best of the Author's knowledge, this is the first example of *multi-modal* person re-identification. While appearance cues are currently the most widely used descriptors for the task of person re-identification, the Chapter provides empirical evidence that recognition performance can be improved by exploiting also anthropometric cues.

It is interesting to point out some possible future research directions in the context of multi-modal re-identification using RGB-D cameras. First, a wider range of anthropometric cues could be used, to further improve recognition capability. In particular, since not all possible anthropometric measures of interest can be extracted from a given image or frame (e.g., because of the pose of the individual), a framework that takes into account missing modalities should be developed, to make it possible to exploit any subset of available modalities. A second research direction is to extend the range of modalities, beyond clothing appearance and anthropometric measures. Skeleton-based gait [59] could be a further cue to exploit, based on the skeleton extraction capabilities of the Kinect SDK. The combination with remote face recognition techniques [99] could also help increasing recognition performance. Possibly, the use of several, distinct modalities could boost recognition performance of re-identification systems towards the one of systems based on strong biometrics.

Chapter 7

Using MCD for appearance-based people search

In this Chapter, the MCD-based method for performing the task of "appearance-based people search" with any existing person re-identification descriptor, presented in Sect. 3.4, is implemented and evaluated. It is worth to recall that the task at hand consists of finding, among a set of images of individuals, the ones relevant to a *textual* query describing clothing appearance of an individual of interest. Therefore, while it shares a lot of commonalities with person re-identification, it nevertheless differs from it, as in person re-identification the query is an *image* of the person of interest instead of a semantic description of his/her clothing.

The method of Sect. 3.4 is applied to two descriptors, and experimentally tested on a novel data set, consisting in a set of images taken from the VIPeR data set, labelled with respect to a predefined set of clothing characteristics (e.g., "red shirt", "short sleeves").

Details on the appearance descriptors used and on the application of MCD are given in Sect. 7.1. Then the experimental evaluation is proposed in Sect. 7.2. Finally, conclusions and possible directions of future research are provided in Sect. 7.3.

7.1 Implementation

The people search approach Sect. 3.4 has been evaluated using two different descriptors for person re-identification. The first descriptor is MCMimpl (Sect. 4.2). It subdivides body into torso and legs, and represents each part with the HSV colour histograms of a bag of randomly extracted 300 image patches. The second is the SDALF descriptor proposed in [45], which has been already described in Sect. 6.2. A variation of the first descriptor has been also tested: it uses a pictorial structure (Sect. 2.1) to subdivide body into nine parts: arms and legs (upper and lower, left and right), and torso. The corresponding implementations of a people search method are denoted respectively as MCD_1, MCD_2 and MCD_3.

All the above descriptors enable queries related to clothing colour. MCD_1 and MCD_2 should permit queries related to upper or lower body, like "white upper body garment". MCD_3 should also enable more specific queries, like "short sleeves", that may be distinguished by the presence of skin-like colour in lower arms. Finally, the RHSP feature used in

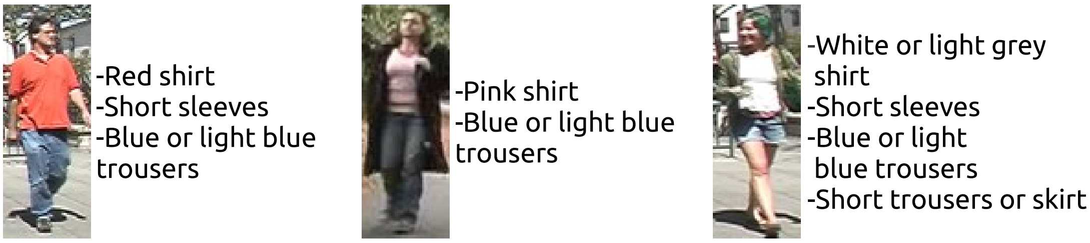

-Red shirt
-Short sleeves
-Blue or light blue trousers
-Pink shirt
-Blue or light blue trousers
-White or light grey shirt
-Short sleeves
-Blue or light blue trousers
-Short trousers or skirt

Figure 7.1: Example images from the VIPeR-Tagged data set.

MCD_2 should should in principle enable queries related to textures, like "checked trousers".

Prototypes are obtained by the a two stage clustering scheme as in Sect. 4.2. In MCD_3, for each body part three different sets of prototypes were created, one for each kind of local features. In the experiments different numbers of prototypes for each body part have been considered, ranging from 5 to 300. The k-th Hausdorff distance was used to compute dissimilarities, with $k = 10$.

7.2 Experimental evaluation

This Section proposes an experimental evaluation of the three people search methods of Sect. 7.1. Sect. 7.2.1 describes the data set; Sect. 7.2.2 explains the experimental set-up. Finally, Sect. 7.2.3 reports the results.

7.2.1 Data set

Experiments have been carried out using a subset of image taken from the VIPeR data set [57], which has been already described in Sect. 4.3.1. Fourteen basic queries related to the colour of the upper and lower body parts, and to the presence of short sleeves/trousers/skirts, have been defined. They are reported in Table 7.1, where the corresponding number of relevant images is also shown. These basic queries have been chosen by considering clothing characteristics that:

1) are detectable to the considered descriptors, and

2) are present in several images of the VIPeR data set, to allow for the construction of a training set of a sufficient size to build the corresponding detectors.

For constructing the training sets, a subset of 512 images from the VIPeR data set was labelled according to the presence of each basic query. This subset of images with the associated labels is denoted in the following as *VIPeR-Tagged*. Some examples are shown in Fig. 7.1

7.2.2 Experimental setup

The retrieval performance of the proposed people search approach on each basic query, for each considered descriptor, has been evaluated by means of the precision-recall (P-R) curve[1]. First, MCD prototypes have been selected from the whole VIPER-Tagged data set,

[1] Precision is the ratio between the number of images correctly labelled as relevant, and the total number of images labelled as relevant. Recall is the ratio between the number of images correctly labelled as relevant,

Class	Cardinality	Class	Cardinality
red shirt	51	green shirt	34
blue/light blue shirt	34	short sleeves	220
pink shirt	35	red trousers/skirt	16
white/light gray shirt	140	black trousers/skirt	12
black shirt	156	white/light gray trousers/skirt	81
orange shirt	10	blue/light blue trousers/skirt	175
violet shirt	18	short trousers/skirt	82

Table 7.1: Labels used to tag the VIPeR-Tagged data set, and corresponding number of positive samples.

using the two-stage clustering scheme of Sect. 4.2.1. Note that prototype creation is unsupervised, as the labels denoting presence/absence of the basic queries are not used in the clustering procedure.

For each basic query, the VIPeR-Tagged has been subdivided into a training and a testing sets of equal size (256 images each), using using a stratified sampling approach to preserve the ratio between relevant and non-relevant images to that class. Then, a statistical classifier has been trained on training images to implement a detector for each basic query. An SVM classifier with linear kernel [31] has been used to this aim. Finally, for each basic query, the P-R curve has been evaluated on testing images, by varying the SVM decision threshold. This procedure has been repeated ten times, and the resulting P-R curves have been averaged to obtain the final results, which are presented in the next Section.

7.2.3 Results

The performance on each basic query is summarised in Table 7.2, in terms of the corresponding average break-even point (BEP), which is the point of the P-R curve whose precision equals recall. The best performance for each basic query is highlighted in bold. As reference, Fig. 7.2 reports four representative examples of the average P-R curves attained. An example of the ten top-ranked images for two basic queries is also shown in Fig. 7.3.

and the total number of relevant images. Precision and recall depend on the parameters that govern the final decision between *relevant* and *not relevant*, e.g. the score threshold. By varying such parameters, it is possible to plot a curve of precision and corresponding recall values.

Class	MCD_1	MCD_2	MCD_3	Class	MCD_1	MCD_2	MCD_3
red shirt	**0.845**	0.780	0.792	green shirt	**0.687**	0.594	0.619
blue/light blue shirt	**0.645**	0.523	0.494	short sleeves	0.631	0.608	**0.643**
pink shirt	0.534	**0.578**	0.461	red trousers/skirt	0.713	0.638	**0.916**
white/light gray shirt	**0.771**	0.736	0.758	black trousers/skirt	0.683	0.607	**0.711**
black shirt	0.728	0.705	**0.736**	white trousers/skirt	**0.758**	0.639	0.635
orange shirt	**0.689**	0.580	0.463	blue trousers/skirt	**0.641**	0.622	0.620
violet shirt	0.422	0.235	**0.433**	short trousers/skirt	0.416	0.393	**0.557**

Table 7.2: Average break-even point attained using the people search methods implemented from the three considered descriptors.

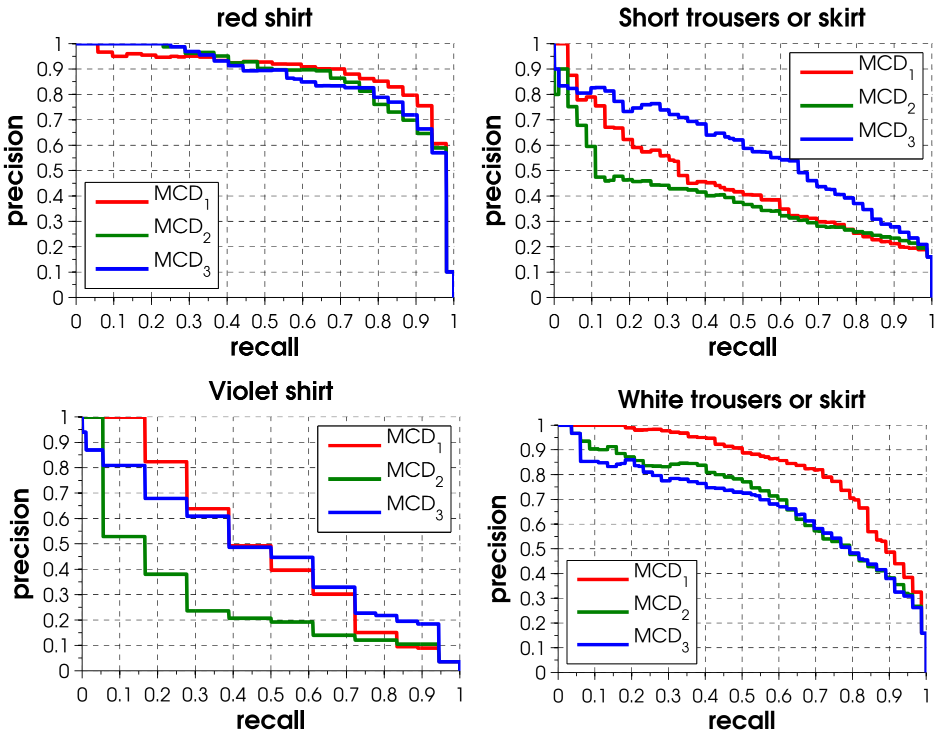


Figure 7.2: Precision-Recall curves attained by the three people search methods on the VIPeR-Tagged data set, for four selected basic queries.

The proposed MCD-based method for appearance-based people search attained a rather good performance with all descriptors, for almost all basic queries. The best performance has been attained on basic queries related to the colours red, white and black (see Table 7.2). The most likely reason is that such colours are well separated in the HSV space, which is used by all the considered descriptors. As stated in Sect. 7.1, MCD$_3$ was likely to attain the best performance on basic queries related to the presence of skin on lower arms and legs, namely "short sleeves" and "short trousers/skirt" (see Fig. 7.2, bottom-left plot), due to its more refined body subdivision. Nevertheless, also MCD$_1$ and MCD$_2$ attained a good performance on these classes. The reason is that, although MCD$_1$ and MCD$_2$ can not distinguish between lower and upper arms or legs, they are nevertheless able to detect skin-like colour in the whole arms or legs, which is strongly related to such basic queries.

As a further investigation, the performance, in terms of average BEP, has been evaluated with respect to the number of prototypes K for each body part, in order to assess the influence of this important parameter of MCD. Results are shown in Fig. 7.4 for MCD$_3$; the other methods show a similar behaviour. As can be observed,the performance initially grows as K increases, then reaches a nearly stable value around $K = 200$ (for MCD$_3$; for the other methods, this value is respectively 100 and 200 for MCD$_1$ and MCD$_2$), depending on the basic query. This behaviour can be easily explained: once the number of prototypes is enough so that most of the distinctive visual characteristics have been captured by different clusters, increasing the number of prototypes has mainly the effect of splitting some of the previous clusters into two or more similar ones. Consequently, no further information is embedded in the new prototypes. Note that the results reported in Table 7.2 and Fig. 7.2 have been attained for $N_m = 200$ (for MCD$_1$) and $N_m = 100$ (for MCD$_2$ and MCD$_3$).

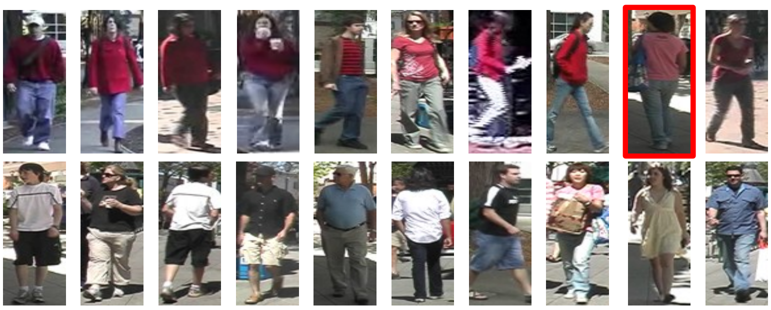


Figure 7.3: The top ten images retrieved by MCD_1, for the "red shirt" (top) and "short sleeves" (bottom) queries, sorted from left to right for decreasing values of the relevance score provided by the detector (classifier). Note that only one non-relevant image is present, highlighted in red.

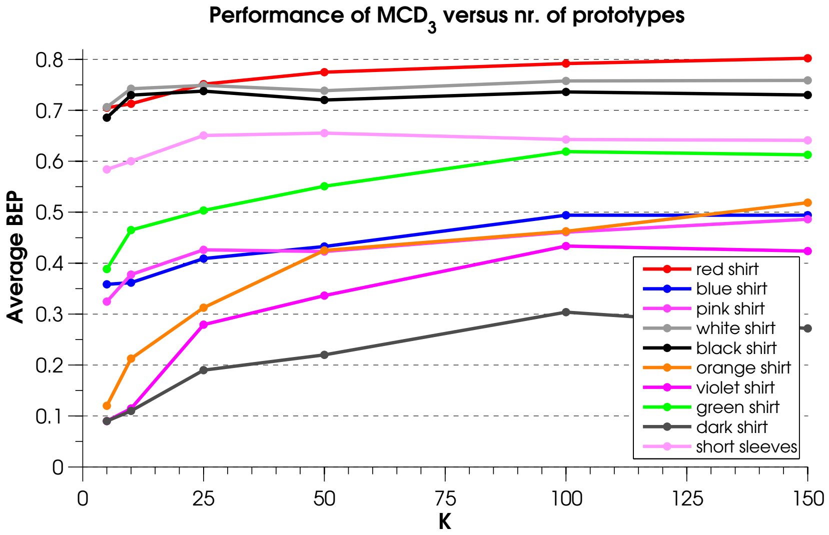


Figure 7.4: Performance of MCD_3, in terms of average BEP, versus the number of prototypes per part K. Note that the other methods, MCD_1 and MCD_2, show a similar behaviour.

7.3 Conclusions

The scope of this preliminary analysis was to experimentally evaluate the general approach of Sect. 3.4 to implement the task of *appearance-based people search*, using the same kind of descriptors used in most existing person re-identification systems.

The approach attained promising results with three different appearance descriptors, on a novel benchmark data set consisting in images taken from the VIPeR data set manually tagged with respect to a set of difference clothing characteristics. An interesting direction of further research is to extend the approach to deal with video sequences. To this aim, pedestrian detection and tracking functionalities that should be deployed as part of a person re-identification system, could be exploited. In this case, a bag of dissimilarity vectors coming from different frames would be available for each person, instead of a single one. A Multiple Instance Learning approach [37] could then be used to train the detectors.

Chapter 8

People search on multimedia data

In the previous Chapters, the reader has been introduced to two tasks for intelligent video-surveillance, namely *person re-identification* and *appearance-based people search*; a general dissimilarity-based framework that can be used to perform these task has been presented and experimentally evaluated in a variety of set-ups (including the combined use of multiple soft biometric modalities to build person descriptors).

Indeed, person re-identification and people search share a lot of commonalities, despite being different tasks. In particular, it is worth to note that both tasks can be seen as *retrieval* problems, where the system returns to the operator a ranked list of results (images or video footages showing people previously seen by the camera network) based on their relevance to a *query*. Basically, from this viewpoint the only difference between these two tasks is in the kind of query given to the system. In the case of person re-identification, the query is an image, or a video-sequence, containing the person of interest; in the case of people search, the query is a *semantic* description of the person of interest's clothing, and needs first to be interpreted by the system.

In this Chapter, the commonalities of the two tasks above are further developed, to formalise a general model of a novel category of retrieval tasks, *people search on multimedia data*. It embraces person re-identification, people search, and number of other possible tasks. It is useful to give first an informal definition of what it is intended here for "people search on multimedia data" (from here on, simply referred to as PSM). Basically, PSM can be described as

> *"the task of retrieving individuals, seen by a network of sensors of any kind, and described using any combination of biometric cues extracted from such sensors, that match a certain criterion (or any combination of criteria) related to these cues".*

The attentive reader should have seen in this definition a direct resemblance to Information Retrieval (IR)[91]. IR is the activity of obtaining documents (texts, images, multimedia information) relevant to a given information need (the query), from a collection of documents. The relationship between IR and PSM grounds the following discussion.

The Chapter is organised in four Sections. Sect. 8.1 formally defines PSM from the classical model or IR. Sect. 8.2 then provides some examples of PSM tasks. Sect. 8.3 suggests MCD as a possible unified framework to perform PSM tasks. Finally, Sect. 8.4 sums up and concludes the Chapter.

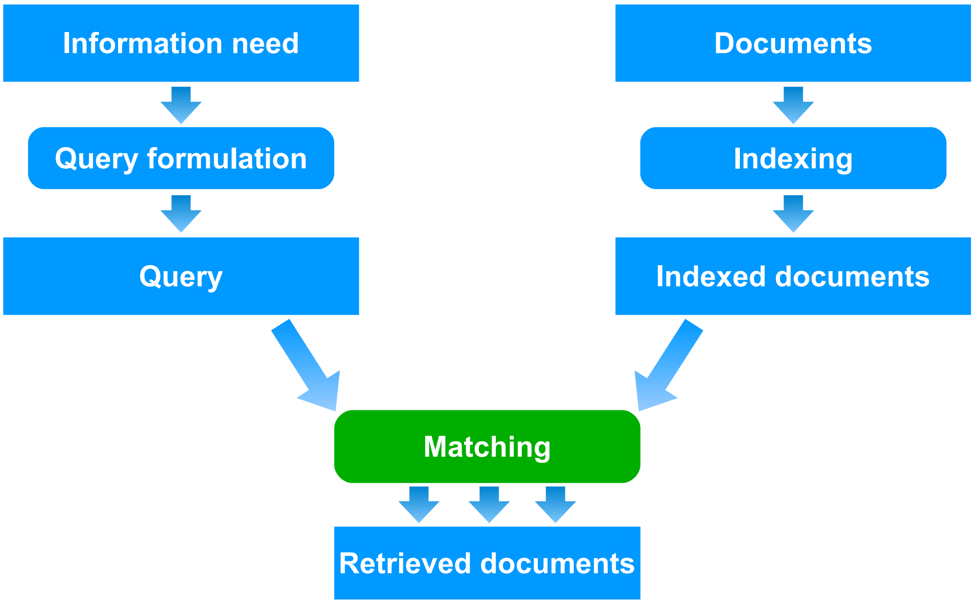


Figure 8.1: Classical Information Retrieval scheme (see the text for details).

8.1 From Information Retrieval to People Search on Multimedia data

Consider the classical model of IR [67], which is graphically shown in Fig. 8.1. Documents where IR shall be performed are at first *indexed*, i.e. represented in some way in a data base. Indexing of a document may consist e.g. of extracting a set of features from it, or associating (either manually or automatically) a set of keywords. Indexing can take place off-line, once for all, on a given fixed set of documents, or on-line when the set of documents is not fixed (every time a new document is added to the set, it is indexed on-the-fly).

The process of IR starts from an *information need* of the user of the IR system. This need has to be formulated into a *query*. Examples of queries are: an image containing the object of interest (e.g. Content-based Image Retrieval), a set of keywords or tags (e.g. textual documents retrieval), an image of a face (e.g. face recognition). Once a query has been formulated, the *matching* phase compares each indexed document to the query according to a certain *matching criterion*, to obtain a scalar value measuring the degree of *relevance* of the document to the given query. For example, in Content-based Image Retrieval the criterion may be a distance measure between the feature vectors (e.g., a colour histogram) of the query image and of each indexed image. Note that that the matching criterion is *fixed* a-priori, and its definition is part of the system design. At the end of the matching phase, all documents are ranked with respect to their degree of *relevance* to the query, and the ranked list is proposed to the user. Possibly, the list of results is truncated so that it contains only the N top-most ranked documents, or those that exhibit a relevance higher than a given threshold.

It can be noted that this IR scheme can directly frame the person re-identification task: the query is an image of the person of interest, and a ranked list of the previously seen (and indexed) templates. The indexing procedure corresponds to descriptor creation. People search can also fit to this scheme. In this case, during matching phase the detectors cor-

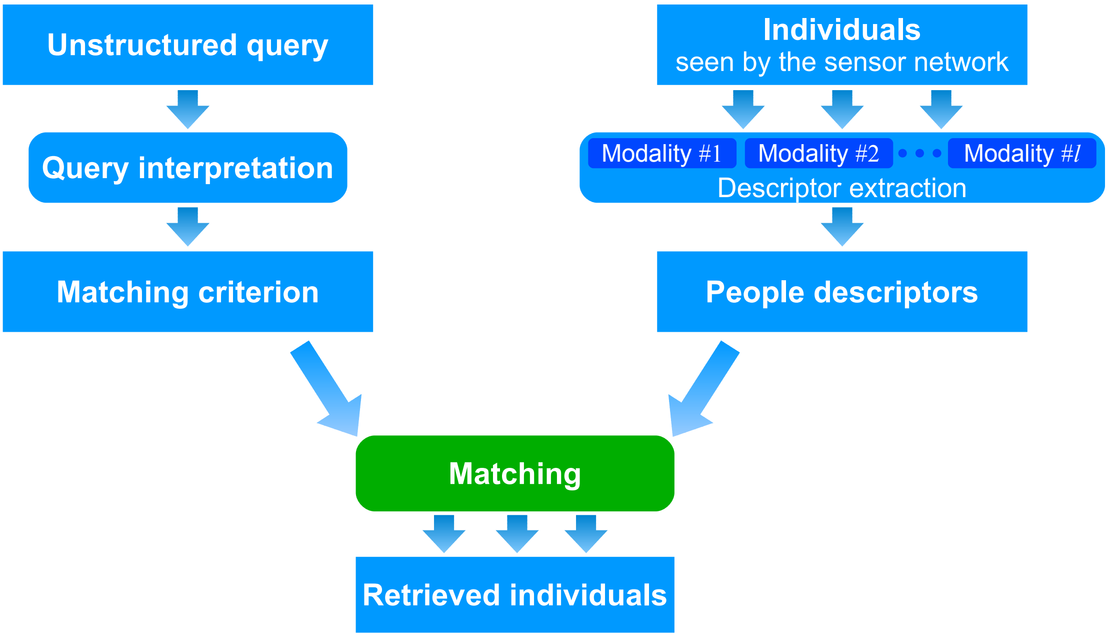


Figure 8.2: People Search on Multimedia data scheme (see the text for details).

responding to the basic queries that compose the query must be run[1]. The final matching score, necessary to produce a relevance ranking, can be any combination of the scores of the basic detectors (e.g. the product of the scores, or the average score).

Both person re-identification and people search are therefore IR problems. Based on this, and building on the scheme of Fig. 8.1, a possible model for PSM, which inherits from IR, can be defined. The first, obvious step is to consider a set of *people* in place of a set of *documents*. Generally speaking, each person can be seen and described using different modalities (e.g., appearance, anthropometry, but also face, speech and/or other biometrics where available) depending on the sensors deployed and on the application scenario. The indexing phase in IR is therefore substituted with the action of extracting a set of *descriptors* from each modality. In other words, each person is represented in the data base as a collection of descriptors. Note that, as a surveillance network is expected to operate continuously, the data base of people descriptors is constantly updated by newly seen individuals.

Concerning the matching phase, as pointed out above, in a typical IR system the corresponding match criterion is fixed, defined once for all at system design, and the query must be formulated accordingly. An useful functionality video-surveillance search systems would be the ability to interpret an user query formulated in ways that are not directly related to the descriptors stored in the data base, e.g. in natural language. This can be useful, for instance, to directly use a textual description of a person, given by a witness, as query, instead of decomposing such description into basic queries as happens in people search (Sect. 3.4). To incorporate this possibility in PSM, a *query interpretation* phase is added, that shall generate the matching criterion that will be used during matching phase. Such phase takes care of translating a unstructured query[2], not directly related to the individuals' representation

[1] Actually, the detectors can be also launched during indexing, and the corresponding outputs can be stored in the database in place of the descriptors.

[2] Here the term "unstructured" refers to queries that do not have a clear, plain, "easy-for-a-computer" structure.

used, into a quantitative criterion on the multi-modal data stored in the data base[3].

The final scheme of the PSM model is shown in Fig. 8.2. Formally, PSM can be described as follows. Let

$$\mathscr{I} = \left\{ I_1 \dots I_n \right\} \tag{8.1}$$

be the set of n individuals seen by the sensors network. The step of descriptor extraction is carried out for each modality, so that each individual is associated to a set of descriptors, one for each modality:

$$I = \left\{ M_{I,1} \dots M_{I,l} \right\} \tag{8.2}$$

where l is the number of modalities.

These descriptor are stored in the data base. Given an unstructured query Q formulated by the user, the process of *query interpretation* takes Q as input an outputs a match criterion in the form of a *membership function* $f_Q(I) = F_Q\left(M_{I,1} \dots M_{I,l}\right)$

$$f_Q : \mathscr{M}_1 \times \mathscr{M}_2 \times \dots \times \mathscr{M}_l \to [0,1] \subset \mathscr{R} \tag{8.3}$$

where $\mathscr{M}_1, \mathscr{M}_2, \dots, \mathscr{M}_l$ are the spaces associated to each modality. The membership function f_Q associates to an individual I a real number in $[0,1]$, indicating the degree of relevance (*score*) of I with respect to the information need codified by the query Q. The *matching* phase applies f_Q to each individual in $\mathscr{I}$. Based on the resulting scores, the elements of $\mathscr{I}$ are ranked and presented to the user.

Similarly to IR, the proposed scheme for PSM can be enriched by adding *relevance feedback* [112, 115], i.e. refining the matching criterion according to an indication of the user that some of the retrieved individuals are truly positive (or negative) with respect to the original information need.

8.1.1 Query interpretation

An important difference between PSM and the classical IR scheme is the presence of a *query interpretation* phase, which in PSM generates the matching criterion. In IR, such criterion is fixed and depends on the task. For example, in Content-based Image Retrieval (CBIR), where the task is to retrieve images similar, in content, to a given one, the matching criterion is a measure of similarity between images[4]. In PSM, the criterion may be changed depending on the information need, to enable unstructured queries like a query given in natural language. For example, consider the natural language query "*white person with a black t-shirt and checked blue shorts, wearing a hat and a pair of sunglasses*". During query interpretation, the system may first find the constitutive semantic concepts that build the query, e.g. "*white person*", "*black shirt*" "*short sleeves*", "*sunglasses*" etc., using Natural Language Processing, then build a membership function that is a combination of the outputs of the detectors associated to these concepts[5].

[3] E.g., a semantic engine able to interpret natural language can be used if the non structured query is a textual description of the person (a task often referred to as Natural Language Processing [69]). This can be done reliably if the application domain and the word ambiguities are limited, such as the case of the description of a person's clothings.

[4] Person re-identification can be seen as a CBIR problem.

[5] Note that such a system may not be equipped with a detector of all the semantic concepts that constitute the query. If not supported concepts are found in the query, they can be signalled to the user, and discarded when building the membership function.

8.2 Example PSM tasks

Building on the model of PSM of Fig. 8.2, various concrete tasks can be defined that can be useful for video-surveillance operators.

Person re-identification, for example, is a PSM task where the the query is an image or a video track showing a person, and the matching criterion (a distance measure between frames or between sets of frames) is fixed. Appearance-based people search is also a PSM task. The query is, in this case, a set of basic queries combined through Boolean operators. The Query Interpretation phase returns a membership function which is a combination of basic detectors. As said in the previous Section, the use of Natural Image Processing could enable a more powerful people search system, able to accept queries formulated in natural language that can be automatically segmented into basic queries. If other cues are extracted from the sensors network in addition to the clothing appearance (e.g., anthropometry), the query can also be related to these cues. E.g., "person with a white shirt, black trousers, *and about 1.80 mt. tall*". Finally, it is easy to add to the Query Interpretation phase the support for *contextual* constraints, e.g., spatio-temporal ones to limit the search to a certain time span or to a subset of the sensor network.

Other PSM tasks can be enumerated that can be of help for surveillance operators and investigators. An example is the retrieval of actions and events of interest, e.g. a person running, jumping, or getting in or out a car in a certain time span [109]. An important difference with respect to the above tasks is that human actions (or events) are *sequences* of configurations of the body, instead of a static characteristic. From a practical viewpoint, this means that the descriptors stored in the database must encode such sequentiality of configurations in a proper way, e.g. a sequence of silhouettes.

Indeed, PSM can be useful also for applications not directly connected to security needs. For instance, fashion-related tasks can be envisaged, like retrieving and counting all the people seen in a shopping centre that wear e.g. a particular kind of jacket, to generate statistics on fashion trends. Since the quality of images taken by the network must be high to enable such task, to do so camera sensors could be mounted on the top of shop windows to have a good capture of people passing by.

8.3 MCD and PSM

In Sect. 3.4, it has been shown that the dissimilarities to MCD prototypes, which represent low level, local or global characteristics of the body or of body parts, can be used as features to train detectors of basic clothing appearance characteristics like "red shirt", "short sleeves" and so on. The motivation of such use of dissimilarities relies on the intuition that certain prototypes (or certain combinations of them) may encode high level concepts. Although this is particularly evident in the case of clothing appearance (see Fig. 3.5 for a clarifying toy example), the assertion that prototypes can be related to high level concepts, and that such relations can be learned using statistical classifiers, is likely to be true also when using domains and modalities different than the appearance, even mixed with it. E.g., the high-level concept "tall person" can be encoded by a low dissimilarity to prototypes of anthropometric measurements corresponding to tall people, and in turn to high dissimilarities to anthropometric prototypes of short people.

This fact motivates the use of MCD as a possible underlying framework for performing

PSM tasks. With respect to the scheme in Fig. 8.1, during Descriptor Extraction individuals are represented as dissimilarity vectors to prototypes encoding characteristics seen with different modalities, and the Criteria are functions of the corresponding dissimilarities (as the spaces $\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_l$ in Eq. (8.3) are dissimilarity spaces). Such use of MCD should ensure a compact individuals' representation to any PSM task. More importantly, the same descriptors and extraction pipeline could be used for many tasks, having only to deal with a proper design of the Criterion formulation phase.

8.4 Conclusions

Relying on the commonalities between person re-identification and people search, and their relationship with Information Retrieval, in this Chapter a possible formulation of the general problem of People Search on Multimedia data (PSM) has been proposed. PSM embraces re-identification and people search, as well as a variety of other useful task for video-surveillance and, possibly, for other application domains.

Along with a formal definition and schematisation of PSM, in Sect. 8.3 the use of MCD as an underlying framework for PSM tasks has been envisaged. This can be an useful hint for various possible future works, some of which will be proposed in Chapter 9.

Chapter 9

Discussion and conclusions

This thesis work presented a contribution to the literature about Intelligent Video Surveillance, a topic that is attracting much interest from researchers and industries due to a continuously growing demand of security and safety inside our present society. In particular, the thesis addressed two tasks, *person re-identification* and *appearance-based people search*, that can provide useful tools for video-surveillance operators and forensic investigators. The latter task is an original contribution of this thesis.

This conclusive Chapter closes the thesis. First, the major contributions of this work are stated in Sect. 9.1. Then the thesis is critically analysed and compared with the present state of the art in Sect. 9.2. Finally, Sect. 9.3 provides future research directions to enrich and extend the present work.

9.1 Contributions of this thesis

The main contribution of this thesis work is a novel framework to construct descriptors of the human appearance for Intelligent Video Surveillance tasks, that is based on *dissimilarity representations*. The framework, called Multiple Component Dissimilarity (MCD), starts from the original dissimilarity paradigm and extends it, in order to deal with objects decomposable in multiple parts and with localised characteristics, as the human body.

MCD has been applied to two Intelligent Video Surveillance tasks, namely person re-identification and appearance-based people search, described above. With respect to these tasks, there are three main advantages of dissimilarity representations for describing persons:

- First, MCD descriptors are compact, and can represent a person using a small vector of real values (dissimilarities). Thanks to this, MCD can drastically reduce computational complexity, specially of the *matching* phase of person re-identification methods.
- Second, MCD builds upon a totally generic formulation of the underlying low-level representation. Therefore, it can be used to combine different descriptors, even if they are heterogeneous (in terms of the model used and/or the features used). Descriptors can even come from different modalities, enabling e.g. *multi-modal* person re-identification, in cases where the clothing appearance is not the only cue available.

- Third, it provides a natural and effective way to learn high-level concepts from low-level representations. This directly enables the task of appearance-based people search described above.

MCD has been exploited in this thesis to achieve several results:

- a method to speed up any existing person re-identification method, which exploits MCD descriptors' compactness to reduce computational needs;
- a state-of-the-art re-identification method, that uses a combination of different kinds of appearance features obtained through the use of MCD;
- a way to combine different descriptors (even heterogeneous and/or coming from different modalities), into a single, compact one, based on MCD;
- building on the last point, a method to perform person re-identification based on two modalities, namely clothing appearance and anthropometry, the latter extracted using RGB-D cameras; to the Author's best knowledge, this is first example of multi-modal person re-identification presented in literature;
- a method that uses MCD to perform the novel task of "appearance-based people search", by learning high level concepts from dissimilarity representations obtained through MCD.

Apart from the above main achievements, this thesis work also provides two important contributions:

- a novel data set for assessing multi-modal person re-identification methods that exploit RGB-D information, made up of RGB and Depth video-sequences showing individuals in different poses and locations, under different illumination conditions;
- a possible formulation of a generalisation of the tasks of people search and person re-identification, that is named "people search on multi-media data".

9.2 Critical analysis

The aim of this Section is to critically analyse MCD and the other major thesis contributions. This Section is subdivided in three parts: in the first part (Sect. 9.2.1), the MCD framework is the subject of the analysis, to the aim of correctly position ot with respect to the state-of-the-art. In the second part (Sect. 9.2.2) the implementations of MCD to carry out person re-identification and people search are analysed compared the rest of the literature. In the third part (Sect. 9.2.3) a more general insight on person re-identification is given, focusing in particular on open problems and unexplored aspects.

9.2.1 Position of MCD with respect to the state-of-the-art

It is important first to correctly position MCD with respect to the field (Intelligent Video Surveillance). In particular, it is worth to point out the role of MCD: rather than being a specific method to accomplish person re-identification or other Intelligent Video Surveillance tasks, it is meant as a general, novel framework to tackle such tasks, that carries three main advantages as listed above:

(i) compactness (which can help in obtaining computationally inexpensive methods for person re-identification),

(ii) independence to the underlying representation (which can enable fusion of multiple descriptors and modalities),

(iii) possibility of a straightforward implementation of appearance-based people search.

To the best of the Author's knowledge, MCD has no direct "competitors" in the literature about person re-identification nor other Intelligent Video Surveillance applications, as no similar frameworks have been proposed so far. However, it shares a similar spirit with other kinds of models for representing objects in Computer Vision. In particular, two of them have resemblances with MCD that are worth exploring.

The first one is the Bag of Words (BoW) model for describing images [135], which is widely used in scene classification. In the BoW model, a vocabulary of *visual words* is at first constructed off-line from a design data set of images, by clustering (usually using k-means) the local features coming from all these images. *Visual words* are then defined as the centroids of the clusters. This step is almost the same as prototype construction in MCD (except that in MCD the different sets of prototypes are constructed, one for each body part). An image is then represented as a normalised histogram of the occurrences of each visual word: each local feature of the image is assigned to the closest visual word, and the corresponding word count is increased by one.

If the number of visual words is not to high (usually, it is around one thousand), the size of a BoW descriptor is comparable to that of a MCD descriptor: the advantage of compactness (i) of MCD can be achieved therefore also with the BoW model. The BoW model is in principle also independent to the kind of local features used (ii), although up to now it has not been used to combine different kinds of local features or multi-modal descriptors. Finally, it is in principle able to feed detectors of basic queries, to implement appearance-based people search, as the same connection between prototypes and high level concepts applies to visual words.

However, MCD and BoW models differ substantially in one major aspect: BoW models count the frequency (count of the occurrences) of each visual word inside each sample, while in the dissimilarity paradigm the *degree of similarity* of each visual prototype is considered. Another important difference between the two is that MCD can support both *local* and *global* features, while BoW can only be used with local features. As various clothing appearance descriptors use global features (possibly, in combination with local ones), this fact limits the applicability of BoW models to the tasks that require the representation of humans. It is worth noting, finally, that BoW and dissimilarity-based representations have been compared in [26] in the specific task of object recognition. In particular, it has been shown that dissimilarity representations outperform BoW representations by a good margin.

Another kind of representation that resembles MCD is the Fisher Vector (FV) model, which inherits from BoW [106]. FV main aim was to obtain a compact image representation with more discriminative power than BoW; basically, it extends BOW by encoding also high-order statistics (first and second order). It relies on Fisher kernels [72], a powerful tool whose underlying idea is to represent a signal with a gradient vector derived from a generative probability model. In FV, this idea is applied to the classic BoW model, using a visual vocabulary obtained as in BoW, to define the generative model: a Mixture of Gaussians centred on each visual word. Fisher Vectors allows for higher classification performance than BoW models in image categorisation tasks. Similarly to BoW, also FV is compact (i), and is in principle feature-independent (ii), although this was not demonstrated in the literature. Finally, as it inherits from BoW it could in principle be used to people search as well. The main limitation of FV with respect to MCD is that, similarly as BoW, it can be used only with local features.

The two models mentioned above exhibit commonalities and differences with respect to MCD. It is indeed desirable, for extending the present work, to explore more thoroughly these commonalities, and to better assess them with respect to Intelligent Video Surveillance tasks. They can be a good source of ideas to enrich MCD as well.

9.2.2 Analysis of the proposed methods for person re-identification and people search

The aim of the following analysis is to highlight critical points of the applications of MCD to implement person re-identification and people search methods, described in Chapters 4, 5, 6 and 7.

Considering first the method proposed in Chapter 4 to speed up existing re-identification methods, on the one hand the experiment analysis clearly demonstrated that the low matching time guaranteed by MCD descriptors enables real-time or quasi-real-time person re-identification. On the other hand, recognition performance of MCD descriptors may be worse than that exhibited by the original descriptor. It was shown that the trade-off between computational time and accuracy can be advantageous for certain kind of applications (e.g., the real-time application scenario described in Sect. 4.1). However, the model described in Sect. 4.1) to evaluate such trade-off makes some simplifications that may be questionable. The first one is the assumption that humans have always 100% accuracy. In other words, given two images of persons, an human should always be able to tell if the two images show the same individual or not. While this assumption seems reasonable at a first sight, it does not take into account that the bad quality of images taken by video-surveillance images cameras (e.g. low resolution, blur, over-exposed or under-exposed, etc.) may pose a challenge also for human operators. This fact is supported by an interesting experiment reported by Cheng et al. [28]; they showed that human operators achieved an average first-rank recognition rate of 75% in a task where each query image had to be found among a set of only 45 images of pedestrians. Furthermore, the human operator may experience a loss of attention depending on its psycho-physical conditions, which may reduce its capability to distinguish persons. Another simplification made by the model is that the time t_c required by an human operator to compare two given images of pedestrian (one probe and one template) in more or less constant. However, depending on the images, the operator may spend more time in comparing details (e.g. if the two persons have very similar clothes), or give immediately a response (e.g. if the two persons have completely different clothes). It is also hard to estimate an average t_c that is generally valid, as it strongly depends on how many people look similar in the data set considered. As such, the model proposed in Sect. 4.1 should indeed be extended, by considering cases where the accuracy of the human operator is not 100% and

t_c is not fixed. These limitations provide possible directions to improve the present work. It is worth to say, however, that the same general conclusion (i.e., that a faster method may be more helpful than a slower but better performing one in on-line re-identification scenarios), is likely to be still valid, even considering the two aspects above.

Consider now the re-identification method proposed in Chapter 5. It shows a solid recognition performance compared to the state-of-the-art, and low computational requirements. There are nevertheless some aspects that could be developed further. The first aspect is the use of MCD to to combine descriptors that already can attain themselves state-of-the-art performance (e.g., SDALF with CPS). It would be very interesting to evaluate their complementarity and if they can be used to attain a higher recognition performance. The second aspect is that the performance of the proposed method is still lower than SDALF and CPS, when considering the first ranks. This could be due to the poor performance of the base descriptors, or due to the inability of dissimilarity representations to maximally exploit their complementarity.

Regarding the multi-modal re-identification method of Chapter 6, the main contributions to highlight are (i) that novel RGB-D sensors can be used effectively for re-identification, and (ii) that combining multiple modalities, and more specifically appearance and anthropometry can help in improving performance. Regarding (i), it is nevertheless important to highlight the present limitations of the RGB-D technology used (the Kinect sensor). First, it can be used indoor only, therefore many possible applications of video-surveillance systems such as monitoring of ports, parking lots, streets, etc. are excluded. In addition, the operative range of the Kinect sensor (which is around 6 mt) is enough only for small closed environments such as corridors and small rooms. The first problem is intrinsic of the IR technology, as the Sun interferes with the IR band. It is however worth noting that the use of high power IR LEDs to generate the IR beam can greatly increase robustness to the Sun light, and permit the use of such devices in outdoor environments where there the sensor is not subject to direct Sun illumination, at the expense of a lower resolution of the Depth map (a technique used e.g. in the Panasonic D-IMager EKL3106). The second problem can be overcome using better and/or more powerful IR emitters and receivers. Both problems can be overcome by using a different technology than IR, such as laser range cameras, at the expense of a higher price per device. Regarding (ii), the used of other anthropometric cues should be explored. Also, as some anthropometric measures can be extracted from a given image or frame only in certain conditions (e.g., frontal pose), a framework that takes into account missing modalities should be developed, to make it possible to exploit any subset of available modalities. Furthermore, it is of interest to explore the used of other modalities, beyond clothing appearance and anthropometric measures. E.g., skeleton-based gait [59] or remote face recognition techniques [99].

Finally, considering the method for appearance-based people search described in Chapter 7, it has been shown that it is possible to effectively provide a very useful novel functionality to forensic investigators using an existing person re-identification method. Nevertheless, a lot of work has to be carried out on this topic. Firstly, Chapter 7 explored the use of still images only, while it would be far more interesting to work with video-sequences. Secondly, the performance was assessed with respect to basic queries only (while it would be useful to evaluate retrieval capabilities when using combinations of basic queries), and by considering only color-related concepts (other concepts, for instance related to repeated patterns like “striped shirt” or “checked trousers” should be implemented). Apart from these limitations, that can provide directions for a further development of the method, there is one major problem that must be taken into account in future research, that is, the *ambiguity* of concepts. Specifically, in this context *ambiguity* means that the definitions of certain concepts may be subjective and/or partially overlap. E.g., “pink shirt” may be confused with a “red shirt” and vice-versa depending on personal taste and/or environmental conditions; similarly, dark colours can be confused with “black”, and other similar examples can be made. This problem must be taken into account and a way to address it, even partially, should be envisaged in future research.

9.2.3 Still unexplored aspects of person re-identification

Person re-identification is a relatively young research area. As such, many aspects have still to be explored, and a large amount of work has to be done before re-identification systems can be used widely in real-world scenarios. Perhaps one of the most important issues that emerge from an analysis of the current literature is that all works focus on *performance*, evaluated mostly on the same three or four benchmark data sets (mostly VIPeR and i-LIDS). Although it is fundamental to assess performance in order to compare methods, various equally fundamental aspects have been almost or totally overlooked. Among them:

(i) computational complexity;

(ii) practical limitations;

(iii) ways of interacting with the operator.

Concerning point (i), the computational complexity of many methods is too high to be used in on-line applications (e.g., SDALF and CPS, see Chapter 5). These methods therefore represent a mere academic exercise from the viewpoint of industries and video-surveillance operators. This thesis work tried to address this issue more systematically, however a more thorough analysis of the true requirements of re-identification systems in terms of computational resources should be performed.

Point (ii) refers to the conditions that can reduce or even nullify the performance of person re-identification systems. E.g., partial occlusions, illumination conditions, the presence of a high number of persons seen in one view, etc. Again, these aspects have not been analysed in depth by current works. Indeed, each of them should be thoroughly studied, to assess their practical influence and make operators aware of what re-identification systems can do, and cannot do. Eventually, the aim should be to develop a framework for comparing re-identification systems.

Point (iii) is about the design of interfaces and ways of interaction that facilitate the use of person re-identification systems by operators. It is indeed an aspect less related to research; still, it is an important step towards moving person re-identification from academic articles to practical products and tools.

9.3 Future works

Various directions for future research can be envisaged. Many possible improvements and aspects to further investigate have been pointed out in the previous Section. Apart from them, the present work can constitute the basis for at least four novel research themes, which are described in the following.

Fusion of dissimilarities and attributes to describe persons.

The MCD framework allows for the construction of dissimilarity vectors to describe people. Dissimilarity vectors represent objects as distances to prototypes encoding local or global low-level characteristics. Apart from low-level characteristics, objects (in this case, people) can be also described by, or be associated with, a set of attributes. E.g., attributes relevant for person re-identification tasks would be those related to the clothing, e.g. a person may wear a *coat*, a *bag*, *sunglasses*, a *t-shirt*, and so on. These two sources informations (dissimilarities and attributes) can be conveniently used together to obtain a better description of pedestrians.

The exploitation of the combination of attributes with low-level features has already been proposed by Layne et al. in a recent paper [81] showing promising results. Layne et al. adopted a score-level fusion scheme where matching is performed using low-level features and attributes separately, obtaining two different matching scores, then the final matching score is obtained as a linear combination of the two.

However, instead of fusing these two sources of information at score level, another possibility is to concatenate dissimilarities (that encode low-level information) with a vector encoding the presence/absence of the attributes (binary values, 1 or 0, or real values representing the output of attribute detectors). In a sense, this is similar to what has been done in Chapter 6 to fuse different modalities, except that in this case one modality is the list of attributes.

Doing so, low-level prototypes and high-level attributes are put at the same level: basically, attributes are seen as *prototypes with a semantic meaning*. Preliminary results have shown a potential usefulness of such fusion technique in increasing re-identification performance when dissimilarity vectors are concatenated with the degree of presence of 14 attributes, estimated via the detectors used in Chapter 7. One major issue is that the weights of the attributes in computing the matching score has to be artificially increased in order to balance the uneven proportion between the number of attributes (usually, tens) and the number of prototypes (usually, hundreds or even thousands). Another important issue is that the final re-identification performance strongly depends on the performance of detectors. Despite these issues, this novel way to represent objects, derived from the dissimilarity paradigm, looks promising and demands for further research.

Development of the PSM model.

Chapter 8 presented a first tentative to formalise a general model of a novel category of retrieval tasks, namely *people search on multimedia data* (PSM). This model frames both person re-identification and appearance-based people search, as well as a number of other possible tasks. Among them, the possibility to implement an enhanced people search supporting queries in natural language has been envisaged, using Natural Language Processing. This could be a very interesting functionality that is worth exploring in future work. Also, it would be useful to further develop the PSM model, enhancing it by adapting techniques and concepts derived from the literature on Information Retrieval (e.g., the use of relevance feedback).

Generalisation of MCD to other domains.

In this thesis, the MCD framework has been developed and tailored specifically for the task of representing pedestrians. It is nevertheless worth to note that the multiple parts-multiple components representation underlying MCD is not limited to humans. Indeed, the same representation can be profitably used in other applicative domains. An interesting direction of further research would therefore be to generalise MCD to other Computer Vision domains. More specifically, MCD would be a nice starting point to develop a novel dissimilarity-based paradigm for Computer Vision. In order to do so, it would be important to further develop on the commonalities shared by MCD and the BoW and FV models described in Sect. 9.2.1.

Bibliography

[1] C. Ben Abdelkader and Y. Yacoob. Statistical estimation of human anthropometry from a single uncalibrated image. *Computational Forensics*, 2008. [cited at p. 23]

[2] Yali Amit and Augustine Kong. Graphical templates for model registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):225–236, mar 1996. [cited at p. 13, 14]

[3] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, june 2008. [cited at p. 5]

[4] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1014–1021, 2009. [cited at p. ix, 15, 16, 61]

[5] Ognjen Arandjelovic and Roberto Cipolla. Face recognition from video using the generic shape-illumination manifold. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz, Austria, May 7-13, 2006, Proceedings, Part IV*, volume 3954 of *Lecture Notes in Computer Science*, pages 27–40. Springer, 2006. [cited at p. 23]

[6] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning implicit transfer for person re-identification. In *Proceedings of the European Conference of Computer Vision (ECCV) Workshops, 1st Workshop on Re-Identification (REID)*, pages 381–390, 2012. [cited at p. 12, 17, 20]

[7] Walid Ayedi, Hichem Snoussi, and Mohamed Abid. A fast multi-scale covariance descriptor for object re-identification. *Pattern Recognition Letters*, 33(14):1902–1907, 2012. Special Issue on Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context. [cited at p. 12]

[8] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person re-identification using haar-based and dcd-based signature. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2010. [cited at p. 13, 17, 20, 28]

[9] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 435–440, 2010. [cited at p. 18, 28]

[10] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *Proceedings of the European Conference of Computer Vision (ECCV) Workshops, 1st Workshop on Re-Identification (REID)*, 2012. [cited at p. 2, 6, 23, 34, 63]

[11] Carlos Barrón and Ioannis A. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding*, 81(3):269–284, March 2001. [cited at p. 23]

[12] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computuer Vision and Image Understanding*, 110(3):346–359, June 2008. [cited at p. 20]

[13] Loris Bazzani, Marco Cristani, Alessandro Perina, Michela Farenzena, and Vittorio Murino. Multiple-shot person re-identification by hpe signature. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 1413–1416, Washington, DC, USA, 2010. IEEE Computer Society. [cited at p. 12, 17, 21]

[14] Loris Bazzani, Marco Cristani, Alessandro Perina, and Vittorio Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, 2012. Special Issue on Awards from ICPR 2010. [cited at p. 13, 17, 21]

[15] A. Bedagkar-Gala and Shishir K. Shah. Part-based spatio-temporal model for multi-person re-identification. *Pattern Recognition Letters*, 33(14):1908–1915, October 2012. [cited at p. ix, 15, 17, 20]

[16] Apurva Bedagkar-Gala and Shishir K. Shah. Multiple person re-identification using part based spatio-temporal color appearance model. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1721–1728, nov. 2011. [cited at p. ix, 15, 17, 20]

[17] C. BenAbdelkader and Y. Yacoob. Statistical body height estimation from a single image. In *Proceedings of the 8th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 1–7, 2008. [cited at p. 23]

[18] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2006. [cited at p. 34]

[19] Soma Biswas, Kevin W. Bowyer, and Patrick J. Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2019–2030, 2012. [cited at p. 23]

[20] Henri Bouma, Sander Borsboom, Richard J. M. den Hollander, Sander H. Landsmeer, and Marcel Worring. Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination. *Proceedings SPIE 8359, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense XI*, pages 83590Q–83590Q–10, 2012. [cited at p. 12, 18, 19, 20, 28]

[21] Thierry Bouwmans, Fida El Baf, and Bertrand Vachon. Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. *Recent Patents on Computer Science*, 1(3):219–237, November 2008. [cited at p. 11]

[22] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980. [cited at p. 18]

[23] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovovery*, 2(2):121–167, June 1998. [cited at p. 15]

[24] Yinghao Cai and Matti Pietikäinen. Person re-identification based on global color context. In *Proceedings of the Tenth International Workshop on Visual Surveillance (VS)*, ACCV'10, pages 205–215, Berlin, Heidelberg, 2011. Springer-Verlag. [cited at p. 18, 20]

[25] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, nov. 1986. [cited at p. viii, 14]

[26] Anna Carli, Umberto Castellani, Manuele Bicego, and Vittorio Murino. Dissimilarity-based representation for local parts. In *Proceedings of the 2nd IEEE International Workshop on Cognitive Information Processing (CIP)*, pages 299–303, 2010. [cited at p. 89]

[27] Savvas .A. Chatzichristofis and Yiannis .S. Boutalis. Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 191–196, may 2008. [cited at p. 51]

[28] Dong S. Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 68.1–68.11, 2011. [cited at p. ix, xi, 15, 17, 20, 21, 34, 50, 53, 59, 60, 61, 62, 90]

[29] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, August 1995. [cited at p. 27]

[30] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 24:603–619, 2002. [cited at p. 41]

[31] Nello Cristianini and John Shawe-Taylor. *An introduction to support Vector Machines and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000. [cited at p. 33, 77]

[32] Angela D'angelo and Jean-Luc Dugelay. A statistical approach to culture colors distribution in video sensors. In *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, AZ, United States, 01 2010. [cited at p. 18]

[33] Angela D'angelo and Jean-Luc Dugelay. People re-identification in camera networks based on probabilistic color histograms. In *SPIE 2011, Electronic Imaging Conference on 3D Image Processing (3DIP) and Applications, Vol. 7882, 23-27 January, 2011, San Francisco, CA, USA*, San Francisco, United States, 01 2011. [cited at p. 12, 18, 20]

[34] G. S. Daniels and E. Churchill. The average man? *Technical Note WCRD TN 53-7: Wright-Patterson Air Force Base, OH: Wright Air Force Development Center*, 1952. [cited at p. 23]

[35] I.O. de Oliveira and J.L. de Souza Pio. Object reidentification in multiple cameras system. In *Proceedings of the 4th International Conference on Embedded and Multimedia Computing (EM-Com)*, pages 1–8, 2009. [cited at p. 20, 28]

[36] Yining Deng, B. S. Manjunath, Charles Kenney, Michael S. Moore, Student Member, and Hyundoo Shin. An efficient color representation for image retrieval. *IEEE Transactions on Image Processing*, 10:140–147, 2001. [cited at p. 18]

[37] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997. [cited at p. 79]

[38] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, april 2012. [cited at p. 5, 11]

[39] Gianfranco Doretto, Thomas Sebastian, Peter Tu, and Jens Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2:127–151, 2011. [cited at p. 1, 2, 4]

[40] Yuning Du, Haizhou Ai, and Shihong Lao. Evaluation of color spaces for person re-identification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, Washington, DC, USA, 2012. IEEE Computer Society. [cited at p. 17, 18, 21]

[41] Marie-Pierre Dubuisson and Anil K. Jain. A modified hausdorff distance for object matching. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, ICPR*, volume 1, pages 566 –568 vol.1, oct 1994. [cited at p. 26]

[42] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2. edition, 2001. [cited at p. 7, 15, 16, 25, 26, 27]

[43] D. Duque, H. Santos, and P. Cortez. Prediction of abnormal behaviors for intelligent video surveillance systems. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 362–367, 1 2007-april 5 2007. [cited at p. 1]

[44] Ahmed M. Elgammal, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II*, ECCV '00, pages 751–767, London, UK, UK, 2000. Springer-Verlag. [cited at p. 5]

[45] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2360–2367, 2010. [cited at p. viii, ix, x, xi, 13, 14, 17, 20, 21, 28, 29, 39, 40, 42, 50, 52, 59, 60, 61, 62, 63, 65, 67, 68, 75]

[46] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. [cited at p. 15]

[47] Pedro F. Felzenszwalb. Representation and detection of deformable shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):208–220, feb. 2005. [cited at p. 14]

[48] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005. [cited at p. 15, 16]

[49] Graham Finlayson, Steven Hordley, Gerald Schaefer, and Gui Yun Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38(2):179–190, 2005. [cited at p. 18]

[50] European Forum for Urban Security. Citizens, cities and video surveillance, 2010. [cited at p. 1]

[51] Per-Erik Forssén. Maximally stable colour regions for recognition and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, USA, June 2007. IEEE Computer Society. [cited at p. ix, 20, 21]

[52] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, EuroCOLT '95, pages 23–37, London, UK, UK, 1995. Springer-Verlag. [cited at p. 16, 21]

[53] Andrew C. Gallagher, Andrew C. Blose, and Tsuhan Chen. Jointly estimating demographics and height with a calibrated camera. In *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, pages 1187–1194, 2009. [cited at p. 23]

[54] Graeme Gerrard and Richard Thompson. Two million cameras in the uk. *CCTV Image Magazine*, 42:10–12, Winter 2011. [cited at p. viii, 2]

[55] Niloofar Gheissari, Thomas B. Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1528–1535, 2006. [cited at p. viii, 13, 14, 17, 20, 28]

[56] Afzal Godil, Patrick Grother, and Sandy Ressler. Human identification from body shape. In *Proceedings of the 4th International Conference on 3D Digital Imaging and Modeling (3DIM)*, pages 386–393, 2003. [cited at p. 23]

[57] Douglas Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 41–47, 2007. [cited at p. viii, 2, 6, 33, 42, 76]

[58] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the 10th European Conference on Computer Vision (ECCV)*, pages 262–275, 2008. [cited at p. 12, 17, 19, 20, 21, 28]

[59] Junxia Gu, Xiaoqing Ding, Shengjin Wang, and Youshou Wu. Action and gait recognition from recovered 3-d human joints. *Transaction on System, Man and Cybernetics, Part B*, 40(4):1021–1033, August 2010. [cited at p. 22, 73, 91]

[60] Bahadir K. Gunturk, Aziz Umit Batur, Yucel Altunbasak, Monson H. Hayes III, and Russell M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–606, 2003. [cited at p. 23]

[61] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal on Machine Learning Research*, 3:1157–1182, March 2003. [cited at p. 27]

[62] M. Hahnel, D. Klunder, and K.-F. Kraiss. Color and texture features for person recognition. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks*, volume 1, july 2004. [cited at p. 19]

[63] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Interest points harvesting in video sequences for efficient person identification. In *Proceedings of the 8th International Workshop on Visual Surveillance (VS)*, 2008. [cited at p. 12, 20, 28]

[64] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proceedings of the Second ACM/IEEE International Conference on Distributed Smart Cameras, 2008. ICDSC 2008.*, pages 1–6, sept. 2008. [cited at p. 12, 20]

[65] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analisys and Machine Intelligence*, 28(2):316–322, February 2006. [cited at p. ix, 22]

[66] Pablo H. Hennings-Yeomans, Simon Baker, and B. V. K. Vijaya Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2008. [cited at p. 23]

[67] Djoerd Hiemstra. *Information Retrieval Models*, pages 1–19. John Wiley & Sons, Ltd, 2009. [cited at p. 82]

[68] Martin Hirzer, Peter M. Roth, and Horst Bischof. Person re-identification by efficient impostor-based metric learning. In *Proceedings of the Ninth IEEE International Conference on Advanced Video and Signal-Based Surveillance, (AVSS)*, pages 203–208, 2012. [cited at p. 12, 20, 21]

[69] Nitin Indurkhya and Fred J. Damerau. *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2nd edition, 2010. [cited at p. 84]

[70] M. Isard and J. MacCormick. Bramble: a bayesian multiple-blob tracker. In *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 34–41, 2001. [cited at p. 1, 5]

[71] David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Classification with nonmetric distances: image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583 –600, jun 2000. [cited at p. 26]

[72] Tommi Jakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11 (NIPS)*, pages 487–493, 1998. [cited at p. 89]

[73] Omar Javed, Khurram Shafique, and Mubarak Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 26–33, june 2005. [cited at p. 12, 17]

[74] Kui Jia and Shaogang Gong. Multi-modal tensor face for simultaneous super-resolution and recognition. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV ’05, pages 1683–1690, Washington, DC, USA, 2005. IEEE Computer Society. [cited at p. 23]

[75] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’02, pages 133–142, New York, NY, USA, 2002. ACM. [cited at p. 21]

[76] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey. Stel component analysis: Modeling spatial correlations in image class structure. *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2044–2051, 2009. [cited at p. 40]

[77] Kai Jungling, C. Bodensteiner, and M. Arens. Person re-identification in multi-camera networks. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 55–61, june 2011. [cited at p. 12, 17, 20]

[78] Arif Khan, Jian Zhang, and Yang Wang. Appearance-based re-identification of people in video. In *Proceedings of the 2010 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 357–362, dec. 2010. [cited at p. 12, 17, 19, 20]

[79] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360, oct. 2003. [cited at p. 4]

[80] Microsoft®Kinect™. http://www.microsoft.com/en-us/kinectforwindows/. [cited at p. 2, 6, 23, 34]

[81] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In *Proceedings of 12th European Conference on Computer Vision (ECCV), Workshop and Demonstrations, First Workshop on Re-Identification (REID2012)*, pages 402–412, 2012. [cited at p. 93]

[82] Kual-Zheng Lee. A simple calibration approach to single view height estimation. In *Proceedings of the 9th Conference on Computer and Robot Vision*, pages 161–166, 2012. [cited at p. 23]

[83] Seok-Han Lee, Tae-Eun Kim, and Jong-Soo Choi. A single-view based framework for robust estimation of heights and positions of moving people. In *Digest of Technical Papers of the 2010 International Conference on Consumer Electronics (ICCE)*, pages 503–504, 2010. [cited at p. 23]

[84] B. Li, H. Chang, S. Shan, and X. Chen. Low-resolution face recognition via coupled locality preserving mappings. *IEEE Signal Processing Letters*, 17(1):20–23, January 2010. [cited at p. 23]

[85] Bo Li, Hong Chang, Shiguang Shan, and Xilin Chen. Coupled metric learning for face recognition with degraded images. In *Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning*, ACML '09, pages 220–233, Berlin, Heidelberg, 2009. Springer-Verlag. [cited at p. 23]

[86] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *Proceedings of the European Conference of Computer Vision (ECCV) Workshops, 1st Workshop on Re-Identification (REID)*, 2012. [cited at p. 12, 17, 19]

[87] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. [cited at p. 19]

[88] Bipeng Ma, Yu Su, and Frederic Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Proceedings of the European Conference of Computer Vision (ECCV) Workshops, 1st Workshop on Re-Identification (REID)*, pages 413–422, 2012. [cited at p. 12, 19, 20, 21]

[89] C. Madden and M. Piccardi. Height measurement as a session-based biometric for people matching across disjoint camera views. In *In Image and Vision Computing New Zealand*, page 29, 2005. [cited at p. 23]

[90] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, jun 2001. [cited at p. 12, 19]

[91] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008. [cited at p. 9, 81]

[92] Niki Martinel and Gian Luca Foresti. Multi-signature based person re-identification. *Electronics Letters*, 48(13):765–767, 21 2012. [cited at p. 13, 20]

[93] Niki Martinel and Christian Micheloni. Re-identify people in wide area camera network. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 31–36, june 2012. [cited at p. 13, 20, 28]

[94] Alexis Mignon. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2666–2672, Washington, DC, USA, 2012. IEEE Computer Society. [cited at p. 22]

[95] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transaction on Pattern Analysis and Maching Intelligence*, 27(10):1615–1630, October 2005. [cited at p. 16]

[96] Arthur G. Money and Harry Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, 2008. [cited at p. 1]

[97] Javier R. Movellan. Tutorial on Gabor Filters. *Tutorial paper http://mplab.ucsd.edu/tutorials/pdfs/gabor.pdf*, 2008. [cited at p. 19]

[98] S. P. Neugebauer and P. A. Sallee. New 3d biometric capabilities for human identification at a distance. In *Proceedings of the 2009 Special Operations Forces Industry Conference (SOFIC)*, 2009. [cited at p. 23]

[99] Jie Ni and Rama Chellappa. Evaluation of state-of-the-art algorithms for remote face recognition. In *Proceedings of the 2010 International Conference on Image Processing (ICIP)*, pages 1581–1584, 2010. [cited at p. 22, 73, 91]

[100] W. Niu, L. Jiao, D. Han, and Y.-F. Wang. Real-time multiperson tracking in video surveillance. In *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, and Fourth Pacific Rim Conference on Multimedia*, volume 2, pages 1144–1148, dec. 2003. [cited at p. 1, 5]

[101] D.B. Ober, S.P. Neugebauer, and P.A. Sallee. Training and feature-reduction techniques for human identification using anthropometry. In *Proceedings of the Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, pages 1 –8, Sept. 2010. [cited at p. 23]

[102] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. [cited at p. 20]

[103] Elbieta Pekalska, Robert P. W. Duin, and Pavel Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, February 2006. [cited at p. 26, 27]

[104] Elzbieta Pekalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications*, volume 64 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005. [cited at p. viii, 2, 7, 25, 26]

[105] Elzbieta Pekalska, Pavel Paclik, and Robert P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, March 2002. [cited at p. 26]

[106] Florent Peronnin and Christopher Dance. Aggregating local descriptors into a compact image representation. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. [cited at p. 89]

[107] Massimo Piccardi and Eric Dahai Cheng. Track matching over disjoint camera views based on an incremental major color spectrum histogram. In *Proceedings of the 2005 IEEE International Conference on Video and Signal Based Surveillance (AVSS 05), 15-16 September 2005, Como, Italy*, pages 147–152. IEEE Computer Society, 2005. [cited at p. 12, 18]

[108] Matti Pietikainen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. *Computer Vision Using Local Binary Patterns*. Computational Imaging and Vision. Springer, Dordrecht, 2011. [cited at p. 50]

[109] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, June 2010. [cited at p. 1, 85]

[110] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 21.1–21.10, 2010. [cited at p. 12, 17, 19, 21]

[111] Deva Ramanan. Learning to parse images of articulated bodies. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2006. [cited at p. ix, 15, 16]

[112] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Series in Automatic Computation, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs NJ, 1971. [cited at p. 84]

[113] J.A. Roebuck, K.H.E. Kroemer, and W.G. Thomson. *Engineering anthropometry methods*. Wiley series in human factors. Wiley-Interscience, 1975. [cited at p. 23]

[114] A. Ross and A. K. Jain. Multimodal Biometrics: an overview. In *Proceedings of 12th European Signal Processing Conference*, pages 1221–1224, 2004. [cited at p. 21, 34]

[115] Yong Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transaction on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998. [cited at p. 84]

[116] Riccardo Satta, Giorgio Fumera, Fabio Roli, Marco Cristani, and Vittorio Murino. A multiple component matching framework for person re-identification. In *Proceedings of the 16th International Conference on Image Analysis and Processing (ICIAP)*, volume 2, pages 140–149, 2011. [cited at p. 39, 61]

[117] Cordelia Schmid. Constructing models for content-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 39–45, 2001. [cited at p. 19]

[118] Sumit Shekhar, Vishal M. Patel, and Rama Chellappa. Synthesis-based recognition of low resolution faces. In *Proceedings of the 2011 International Joint Conference on Biometrics*, IJCB '11, pages 1–6, Washington, DC, USA, 2011. IEEE Computer Society. [cited at p. 23]

[119] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, 2011. [cited at p. 23]

[120] Leonid Sigal and Michael J. Black. Predicting 3d people from 2d pictures. In *Proceedings of the IV Conference on Articulated Motion and Deformable Objects (AMDO)*, pages 185–195, 2006. [cited at p. 16]

[121] Thomas Sikora. The mpeg-7 visual standard for content description - an overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):696–702, June 2001. [cited at p. 19, 51]

[122] Sarah V. Stevenage, Mark S. Nixon, and Kate Vince. Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, 13(6):513–526, 1999. [cited at p. 22]

[123] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 103–110, 2012. [cited at p. 23]

[124] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen. Person attribute search for large-area video surveillance. In *Proceedings of the 2011 IEEE International Conference on Technologies for Homeland Security (HST)*, pages 55–61, 2011. [cited at p. 5]

[125] M. Tkalcic and J.F. Tasic. Colour spaces: perceptual, historical and applicational background. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, volume 1, pages 304–308, sept. 2003. [cited at p. 17]

[126] Dung Nghi Truong Cong, Catherine Achard, Louahdi Khoudour, and Lounis Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP)*, pages 179–189, 2009. [cited at p. 12, 18, 19, 20, 28]

[127] Dung Nghi Truong Cong, Louahdi Khoudour, Catherine Achard, Cyril Meurie, and Olivier Lezoray. People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374, August 2010. [cited at p. 12, 18, 19, 20]

[128] Amos Tversky. Features of similarity. 84(2):327–352, 1977. [cited at p. 26]

[129] Koen van de Sande, Theo Gevers, and Cees Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, September 2010. [cited at p. 17, 18]

[130] Daniel Vaquero, Rogerio Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'09)*, 2009. [cited at p. 5]

[131] Chun-hao Wang, Yongjin Wang, and Ling Guan. Event detection and recognition using histogram of oriented gradients and hidden markov models. In *Proceedings of the 8th International Conference on Image Analysis and Recognition - Volume Part I*, ICIAR'11, pages 436–445, Berlin, Heidelberg, 2011. Springer-Verlag. [cited at p. 1]

[132] Jun Wang and Jean-Daniel Zucker. Solving the multiple-instance problem: A lazy learning approach. In *ICML*, 2000. [cited at p. 41]

[133] Yang Wu, Masayuki Mukunoki, Takuya Funatomi, M. Minoh, and Shihong Lao. Optimizing mean reciprocal rank for person re-identification. In *Proceedings of the 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '11, pages 408–413, Washington, DC, USA, 2011. IEEE Computer Society. [cited at p. 13, 17, 21]

[134] Dong Xu, Yi Huang, Zinan Zeng, and Xinxing Xu. Human gait recognition using patch distribution feature and locality-constrained group sparse representation. *IEEE Transactions on Image Processing*, 21(1):316–326, jan. 2012. [cited at p. 22]

[135] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval (MIR)*, pages 197–206, 2007. [cited at p. 30, 89]

[136] Nai-Chung Yang, Wei-Han Chang, Chung-Ming Kuo, and Tsia-Hsing Li. A fast mpeg-7 dominant color extraction with new similarity measure for image retrieval. *Journal of Visual Communication and Image Representation*, 19(2):92–105, February 2008. [cited at p. 13]

[137] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4), December 2006. [cited at p. 1, 5]

[138] Guoying Zhao, Guoyi Liu, Hua Li, and M. Pietikainen. 3d gait recognition using multiple cameras. In *7th International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 529–534, april 2006. [cited at p. 22]

[139] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009. [cited at p. viii, 6, 42, 43]

[140] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 649–656, Washington, DC, USA, 2011. IEEE Computer Society. [cited at p. 12, 17, 19, 21]

List of Publications

Published papers related to the Thesis

Journal papers

- Riccardo Satta, Giorgio Fumera, Fabio Roli, *Fast Person Re-Identification Based on Dissimilarity Representations*, Pattern Recognition Letters, vol. 33, issue 14, pp. 1838–1848, 2012. (Relation to Chapter 3 and 4)

Conference papers

- Riccardo Satta, Giorgio Fumera, Fabio Roli, Marco Cristani, and Vittorio Murino, *A Multiple Component Matching Framework for Person Re-Identification*, in Proc. 16th Int. Conf. on Image Analysis and Processing (ICIAP 2011), Ravenna, Italy, September 2011. (Relation to Chapter 4)

- Riccardo Satta, Giorgio Fumera, and Fabio Roli, *Exploiting Dissimilarity Representations for Person Re-Identification*, in Proc. 1st Int. Workshop on Similarity-Based Pattern Analysis and Recognition (SIMBAD 2011), Venice, Italy, September 2011. (Relation to Chapter 3)

- Riccardo Satta, Giorgio Fumera, and Fabio Roli, *Appearance-based People Recognition by Local Dissimilarity Representations*, in Proc. 14th ACM Workshop on Multimedia and Security (MMSEC 2012), Coventry, United Kingdom, September 2012. (Relation to Chapter 4 and 7)

- Riccardo Satta, Giorgio Fumera, and Fabio Roli, *A General Method for Appearance-based People Search Based on Textual Queries*, in Proc. 12th European Conference on Computer Vision Workshops and Demonstrations, 1st Workshop on Re-Identification (ReID 2012), Florence, Italy, October 2012. (Relation to Chapter 7)

- Riccardo Satta, Federico Pala, Giorgio Fumera, and Fabio Roli, *Real-time Appearance-based Person Re-identification over Multiple Kinect Cameras*, in Proc. 8th International Conference on Computer Vision Theory and Applications (VISAPP 2013), Barcelona, Spain, February 2013. (Relation to Chapter 5)

Submitted papers related to the Thesis

- Riccardo Satta, Federico Pala, Giorgio Fumera, and Fabio Roli, *Multi-modal Person Re-Identification Using RGB-D Cameras* . Submitted to a journal. (Relation to Chapter 5)

Other published papers, not related to the Thesis

Journal papers

- Chang-Tsun Li and Riccardo Satta, *An Empirical Investigation into the Correlation between Vignetting Effect and the Quality of Sensor Pattern Noise*, IET Computer Vision, vol. 6, issue 6, pp. 560–566, 2012.

Conference papers

- Ignazio Pillai, Riccardo Satta, Giorgio Fumera, and Fabio Roli, *Exploiting Depth Information for Indoor-Outdoor Scene Classification*, in Proc. 16th Int. Conf. on Image Analysis and Processing (ICIAP 2011), Ravenna, Italy, September 2011.

- Chang-Tsun Li and Riccardo Satta, *On the Location-Dependent Quality of the Sensor Pattern Noise and Its Implication in Multimedia Forensics*, in Proc. 4th Int. Conf. on Imaging for Crime Detection and Prevention (ICDP 2011), London, United Kingdom, November 2011.

- Claudio Cusano, Riccardo Satta, and Simone Santini,*Unsupervised Classemes*, in Proc. 12th European Conference on Computer Vision Workshops and Demonstrations, Workshop on Information Fusion in Computer Vision for Concept Recognition (IFCVCR 2012), Florence, Italy, October 2012.