Università degli Studi di Cagliari

DOTTORATO DI RICERCA IN
INGEGNERIA ELETTRONICA E INFORMATICA

Ciclo XXVIII

TITOLO TESI
RE-IDENTIFICATION AND SEMANTIC RETRIEVAL OF PEDESTRIANS IN VIDEO SURVEILLANCE SCENARIOS

Settore scientifico disciplinare di afferenza

ING-INF/05 Informatica

Presentata da:	**Federico Pala**
Coordinatore Dottorato	**Fabio Roli**
Tutor	**Giorgio Fumera**

Esame finale anno accademico 2014 – 2015

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

Re-identification and semantic retrieval of pedestrians in video surveillance scenarios

Federico Pala

Advisor: Giorgio Fumera
Curriculum: ING-INF/05 Computer Engineering

XXVIII Cycle
March 2016

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

Re-identification and semantic retrieval of pedestrians in video surveillance scenarios

Federico Pala

Advisor: Giorgio Fumera
Curriculum: ING-INF/05 Computer Engineering

XXVIII Cycle
March 2016

Dedicated to my family

Abstract

Person re-identification consists of recognizing individuals across different sensors of a camera network. Whereas clothing appearance cues are widely used, other modalities could be exploited as additional information sources, like anthropometric measures and gait. In this work we investigate whether the re-identification accuracy of clothing appearance descriptors can be improved by fusing them with anthropometric measures extracted from depth data, using RGB-D sensors, in unconstrained settings. We also propose a dissimilarity-based framework for building and fusing multi-modal descriptors of pedestrian images for re-identification tasks, as an alternative to the widely used score-level fusion. The experimental evaluation is carried out on two data sets including RGB-D data, one of which is a novel, publicly available data set that we acquired using Kinect sensors.

In this dissertation we also consider a related task, named semantic retrieval of pedestrians in video surveillance scenarios, which consists of searching images of individuals using a textual description of clothing appearance as a query, given by a Boolean combination of predefined attributes. This can be useful in applications like forensic video analysis, where the query can be obtained from a eyewitness report. We propose a general method for implementing semantic retrieval as an extension of a given re-identification system that uses any multiple part-multiple component appearance descriptor. Additionally, we investigate on deep learning techniques to improve both the accuracy of attribute detectors and generalization capabilities. Finally, we experimentally evaluate our methods on several benchmark datasets originally built for re-identification tasks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A feature still not present in nowadays video-surveillance systems, consists on recognizing if a pedestrian acquired by a camera network, has already been seen in some other location. Considering a closed-circuit camera network that cover for instance an entire building, a whole neighbourhood or an airport, this would allow for an automatic mapping of subjects, speeding up the work of law enforcement in case of some investigation. A system able to spot a suspect and take over human operators from watching the entire surveillance recordings, is definitely an interesting option, not just for a matter of time. In fact, watching hours of surveillance recordings actually comport loss of attention and concentration with obvious consequences. With the aid of an automatic system, as soon as the officer notice a suspect behaviour, it would be straightforward establishing what the subject was previously doing. This task is known in the computer vision and pattern recognition community as Person Re-Identification.

The main difficulties deriving from considering video surveillance scenarios are related to the low quality and extensive length of the recordings. In order to record an high quantity of data into physical supports, it is necessary to record videos in low resolution and storing them using some compression algorithm. Another, often insurmountable difficulty, is that the pedestrians can give the back to the surveillance cameras, making it impossible to use face recognition algorithms. Therefore, it is mandatory to exploit some weaker biometry, able to be acquired at a far distance. Even if they are not as much discriminant as the strong biometrics used for user authentication, such soft biometrics would be at least able to narrow down the search space around the suspect.

In general, person re-identification has the purpose of finding the video recording frames that show the subject depicted in the query image. Another interesting application consist on querying a video-surveillance system using, instead of an image, a textual description of the clothing appearance of the searched subject (e.g. look for individuals wearing a red t-shirt). In forensic analysis, this can be useful when the suspect is described by an identikit given by an eyewitness. Another possible application is the search of a subject that has been abducted. In fact, relatives and friends are able to give the authorities some description of what they were wearing before getting lost. These data can be used to query a semantic retrieval system that would be surely quicker than several human operators watching maybe city wide camera recordings. This can give some more chances of retrieving the individual. This task is usually called appearance-based people search, or semantic retrieval, and along

with person re-identification is the main topic of this dissertation.

The branch of research that involves all the thesis work, is for the major part dedicated to the study of models able to characterize the appearance of an individual acquired by a camera. Clothing appearance is the main cue used in literature since clothing characteristics do not change significantly if the subject is acquired from a different point of view, or has a different body pose. At the same time it is not straightforward because, depending on different lighting conditions, colors can appear to be different and usually recordings are not white balanced. Moreover, in the case of partial occlusions and non uniform illumination, the algorithms for people detection and segmentation nowadays are far from perfect. Consequentially, we are restricted to analyse scenarios where there are not too many people, for instance in a parking spot or in a not too much crowded mall.

In this thesis, to face the person re-identification task we use a method based on dissimilarity representations [78]. Instead of exploiting a feature based representation of individuals (i.e. in terms of low level characteristics such as color histograms), we evaluate the dissimilarity in respect to some visual prototypes. Such prototypes describe low level local characteristics shared by individuals. For instance, a color distribution tending to red in the upper part, can be taken as a concept to describe a red shirt. Such prototypes can refer to different body parts: to represent a t-shirt we can consider just the upper part while to represent trousers the lower part.

In order to be able to divide the image of an individual in body parts, we need some algorithm able to establish what pixels are related to the torso and to the legs part. In previous works [85, 83] we used a state of the art method based on pictorial structures [2]: a generative model of the human body that along with some strong part detectors, is able to infer from a still image the body part constellation. With the advancement of RGB-D sensors technology, in this thesis we take advantage of devices such as the Microsoft Kinect, to obtain a more fast and accurate segmentation of the pedestrian image. This is explained in chapter 3, where along with clothing appearance we consider also a set of anthropometric measures that can be acquired at a certain distance from the depth camera. The use of this cue well suites the multiple-component multiple-part representation that we introduced in [85] and then embedded into the dissimilarity-based representation in [83].

Apart from providing a discriminant signature of pedestrians, the Multiple Component Dissimilarity (MCD) representation comes with a framework that assist person re-identification, from the extraction of descriptors to the matching procedure. This framework is also naturally suitable for retrieving people by semantic attributes. Once the visual prototypes are extracted from an image dataset, it is possible to train an attribute detector for each characteristic that we need to retrieve. Each detector is a classifier that maps the dissimilarity representation of the individual into the space of attributes, using a set of training images to learn the prediction model. In this way, when a human operator query the system, it will display all the images in the database that have characteristics similar to the prototypes that are affine to the query description. In addiction, it would be possible to use the system in reversal, starting from the video frame and annotating the image with the clothing characteristics of the reported subject. This can be useful to organize and locate images of interest from the video recordings acquired by a surveillance camera network.

Taking advantage of the recent breakthroughs in image tagging [52] given by deep learning algorithms, we go further in chapter 5 by constructing a new appearance model for the pedestrian image, automatically constructed from the raw pixel data. Whereas the similarity in respect to visual prototypes is usually built from hand crafted features, such algorithms

are able to generate better representations that adapt naturally to the video domain and set of attributes at hand. With regard to the multiple part representation, also the human body subdivision can be inferred by deep learning techniques. This has been exploited by [62] using deep decompositional networks, and allows for excellent performances and real time computation. Moreover, the great expressive capacity of convolutional neural networks, consents to investigate on more complex attributes and get better performance in term of accuracy of the attribute detectors.

The contributions of this thesis can be summarized in these five points:

- we exploit the use of anthropometric measures for person re-identification, acquired at a far distance by RGB-D sensors. We fuse this new information with state of the art clothing appearance descriptors increasing their first-rank recognition rate up to 20%;
- in respect to the state of the art, where clothing appearance and anthropometrics modalities are merged using a score level fusion, we explore a different possibility extending our previously proposed Multiple Component Dissimilarity framework. This fusion approach reduces considerably the processing cost of the person re-identification matching phase ;
- we introduce a general method for implementing semantic pedestrian retrieval using the same Multiple Component Dissimilarity descriptor used for person re-identification;
- we propose a method for processing complex semantic queries, obtained by combining basic ones through boolean operators;
- we introduce a new objective function, aimed at optimizing the performance of convolutional neural networks in semantic retrieval of pedestrians, exploiting the co-occurrence of attributes. This leads to better detection performances and generalization capability to new datasets and attributes.

The thesis is structured as follows. In chapter 2 we introduce the main algorithms that we are going to use in the rest of the work, essentially the Multiple Component Dissimilarity framework. In chapter 3 we describe our contribution to person re-identification through the use of RGB-D cameras. Along with technical details of our implementation we explain how we extended the MCD framework to multiple modalities. In chapter 4 we introduce our MCD extension in order to face the problem of pedestrian semantic retrieval using hand crafted features, and in chapter 5 using deep learning techniques. Chapter 6 concludes the thesis with some general consideration on the whole work, and some insights on future research.

Chapter 2

Related work

In this chapter we overview person re-identification methods based on clothing appearance (which is the only cue used in most of the existing methods), and on their combination with other modalities, focusing on anthropometric measures. In particular, section 2.1 address the state of the art, while section 2.2 presents our previous work on the Multiple Component Dissimilarity (MCD) framework for person re-identification. These sections are useful to fully understand the content of chapter 3. We continue in 2.3 with the literature for the semantic retrieval of pedestrians task, that will be addressed in chapters 4 and 5.

2.1 Person re-identification

Person re-identification consists of matching individuals across different, possibly non overlapping views of a camera network [30]. It can enable various applications, like off-line retrieval of video sequences containing an individual of interest whose image is given as a query, and on-line pedestrian tracking (aka *re-acquisition* [43]). Strong biometric traits, like faces, cannot be exploited in typical settings characterized by strong pose variations, partial occlusions, low resolution and unconstrained environment [30] (see Fig. 2.1).

Clothing appearance is the most widely used "soft", session-based cue, since it is relatively easy to extract, and exhibits uniqueness over limited periods of time. The accuracy it can attain is however affected by low inter-class variability (i.e., different individuals wearing similar clothing), especially in scenarios involving a large number of people. Re-

Figure 2.1: Sample images of a video-surveillance setting, taken from the VIPeR [43] and CAVIAR4REID [18] data sets: low image resolution and unconstrained poses do not allow strong biometric traits, like faces, to be exploited.

identification based only on clothing appearance can actually be difficult also for human operators, as pointed out, e.g., in [18]. For this reason some authors proposed to combine it with other modalities, like anthropometric measures [71, 70], gait [51] and thermal data [70], and reported some evidences that the proposed multi-modal systems can outperform systems based on clothing appearance alone.

In this section we overview person re-identification methods based on clothing appearance (which is the only cue used in most of the existing methods), and on their combination with other modalities, focusing on anthropometric measures.

2.1.1 Person re-identification based on clothing appearance

Most of the existing descriptors are based on a multiple part - multiple component (MPMC) representation: they subdivide the body into several parts to deal with its non-rigid nature, and represent each part as a set of components using various kinds of local or global features. SDALF [31] is a paradigmatic example: it subdivides the body into left and right torso and legs, according to its symmetry and anti-symmetry properties. Three kinds of features are extracted from each part: color histograms in the HSV color space; *maximally stable color regions* (MSCR); *recurrent high-structured patches* (RHSP) (see Sect. 3.2.2). To extract MSCR and RHSP, several image patches are randomly sampled, and then clustered to find the most significant ones. In [68] the body is subdivided into head, torso and legs as in [31]. Each part is described using weighted Gaussian color histograms features (to capture the chromatic content of the region around SIFT key-points), pyramid of histograms of orientation gradients (by concatenating the histogram of gradients along edge lines), and Haralick features (to describe textures, based on the gray level co-occurrence matrix). In [3] the body is subdivided into upper and lower parts: each part is represented using the MPEG7 dominant color descriptor, and a learning algorithm is used to find the most discriminative appearance model. Our MCMimpl [85] subdivides the body into torso and legs, randomly extracts from each part rectangular, possibly overlapping patches, and represents them with HSV color histograms; to attain robustness to illumination changes, synthetic patches are generated from the original ones by varying brightness and contrast. In [67] dense color histograms in several colour spaces, and different texture features, are extracted from four body parts (upper and lower torso, upper and lower legs); a nonlinear warp function between features from two cameras is learnt, to deal with large changes in appearance between them, due to different lighting conditions and poses, occlusion, and background clutter. In [99] the body is subdivided into six horizontal strips, and clothing appearance is modelled in terms of color and texture features as a function of the pose; a classifier is trained offline to identify the most discriminative features, and subject-discriminative features are further learnt online. In [59] a spatial pyramid is built by dividing an image into overlapping horizontal stripes of 16 pixels height; color histograms and histograms of oriented gradients are computed for each strip. A ranking method specific to re-identification, based on sparse discriminative classifiers, is also proposed in [59].

Other methods exploit more refined subdivisions. In [4], a body part detector is used to find fifteen non-overlapping square cells, corresponding to “stable regions” of the silhouette, which are represented by a covariance descriptor in terms of color gradients. Color histogram equalization is performed to improve robustness to changing lighting conditions. Descriptor generation and matching are performed through a pyramid matching kernel. The subdivision of [37] is based on decomposable triangulated graphs, and each part is described

by color and shape features. Pictorial structures are used in [18] to detect chest, head, thighs and legs, which are described by HSV histograms and MSCR patches as in [31].

Other approaches treat the body as a whole, and represent it using various kinds of features: Haar-like features [3]; SIFT-like interest points [37, 46, 25]; texture (Schmid and Gabor filters) and color (histograms in different color spaces) [44]; global color descriptors (histograms, spatiograms, color/path-length) [95]; 4D multi color-height histograms and transform-normalized RGB (illumination-invariant) [16]; biologically-inspired features and covariance descriptors capturing shape, location and color information [64] (see Sect. 3.2.2).

The use of RGB-D sensors has recently been proposed, since they enable a more effective foreground/background segmentation and body part localization, with respect to RGB sensors [1, 86]. In [1] a "body print" is extracted from each individual, exploiting the precision of RGB-D sensors in measuring world coordinates corresponding to the pixels of the scene; the body is subdivided into horizontal stripes at different heights, and each stripe is represented as its mean RGB value along a video sequence. In [86] the Kinect SDK is used for extracting the torso and legs body parts, which are then represented using the same descriptor of [83].

2.1.2 Person re-identification using anthropometric measures

Anthropometry is the characterization of individuals through the measurement of physical body features, e.g., height, arm length, and eye-to-eye distance [80], typically taken with respect to landmark points like elbows, hands, knees and feet, which are localized automatically or manually. Their discriminant capability was discussed in a classic study [22], where ten different measures were evaluated over 4,063 individuals.

Although the use of anthropometric measures has already been proposed for personal identity recognition, existing methods do not fit the typical setting of person re-identification, i.e., multiple, usually non-overlapping cameras and unconstrained environments, with large pose variations and non-cooperative users. In some works, anthropometric measures are extracted using costly 3D devices, like 3D laser scanners, and require the user collaboration in a constrained setting [40, 75, 72]. Other methods use an RGB camera with no specific calibration [8, 11]; however, anthropometric measures can be evaluated in [8] up to a scale factor only, and thus cannot be used for comparing individuals in images acquired by different cameras, whereas in [11] images in frontal pose are required, and thirteen body landmarks have to be manually selected. Other methods estimate absolute height values only, but require ad hoc camera calibration [10, 34, 55, 54].

The use of anthropometric measures for re-identification was first proposed in [66], where height was estimated from RGB cameras as a cue for associating tracks of individuals coming from non-overlapping views (this corresponds to the re-acquisition task that is enabled by person re-identification), but ad hoc camera calibration is required. The extraction of anthropometric measures in re-identification settings has recently been made viable by novel RGB-D sensors, which provide a reliable, real-time body pose estimation [90, 93]. For instance, the Kinect SDK[1] provides the absolute position of twenty body joints (see Fig. 2.2). This was exploited in [7] to extract different anthropometric measures from front and rear 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 centre and right shoulder; three other geodesic

[1] Microsoft®Kinect™Software Development Kit, http://www.microsoft.com/en-us/kinectforwindows/

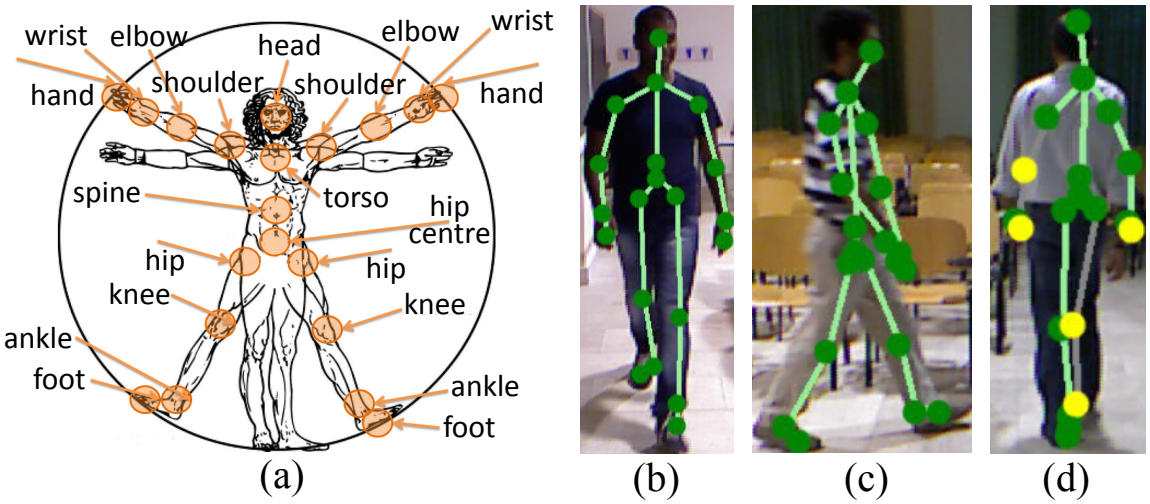

Figure 2.2: (a) The 20 skeletal joints tracked by the Kinect SDK, in the classical Vitruvian Man. (b–d) Depending on the confidence degree on the estimated joint positions, the Kinect SDK distinguishes between *tracked* (in green) and *inferred* (less reliable) points (in yellow). Some joints could also be not tracked, depending on the pose, like the right ankle and foot in (d).

distances are estimated from the 3D mesh of the abdomen, using the Kinect depth map: torso centre to left shoulder, torso centre (located in the abdomen) to left hip, and distance of torso centre to right hip. In [61] a specific setting was considered, in which the cameras are installed on the floor after an entrance door. The proposed anthropometric measures (extracted from a sequence of frames) are the individual's average blob height, area and projected volume to the floor, and blob speed. Two kinds of descriptors have recently been considered in [69], exploiting Kinect sensors: 13 anthropometric measures extracted from the body joints, and a point cloud model of human body. We finally mention that anthropometric measures extracted from RGB-D sensors have been recently proposed also for person recognition tasks; e.g., height, arm and leg length were used in [38] and [39] for gait recognition.

We point out that some of the measures of [7] are difficult or even impossible to extract from unconstrained poses: measures from 3D mesh require near-frontal pose (the abdomen is hidden in rear pose); neck distance to left and right shoulders is hard to compute from lateral pose, even using a depth map, and requires to distinguish between left and right body parts. The measures used in [61] are tailored to the specific setting of top-camera views, instead, and cannot be used in standard re-identification settings. The skeleton measures of [69] are in principle suited to standard settings, and can be extracted from unconstrained poses.

2.1.3 Multi-modal person re-identification

Some authors recently proposed to combine clothing appearance with cues coming from other modalities: anthropometric measures, thermal features and gait. In [70] clothing appearance (RGB histograms from upper and lower body) was combined with thermal features (SURF descriptors), and anthropometric measures (averaged over different frames from the depth map): frontal curve model, encoding the distances from head to neck and neck to torso along the body surface; thoracic geodesic distances; and the length of seven inter-

joint segments connecting different body parts. In [71] clothing appearance (color histogram of upper body and legs) was combined with the subject's height, computed by subtracting the y-values in world-coordinates of the upmost and the lowermost silhouette points. In [51] clothing appearance (color histograms) was combined with gait, described by a spatio-temporal histogram of oriented gradients. In [15] clothing appearance (color histogram of head, torso and legs, and a texture model based on Local Binary Patterns) was combined as in [71] with the height, which was however extracted using only RGB sensors, by converting the detected part part positions in real world coordinates through a specific camera calibration.

We point out that in these works the improvement in re-identification accuracy that can be attained by the additional modalities over clothing appearance has not been clearly evaluated. Moreover, in all the existing methods where multi-modal cues are used, the different modalities are combined through score-level fusion techniques. A different fusion technique between different descriptors was proposed in [33], although it was evaluated only on clothing appearance cues: it exploits a multi-view, semi-supervised learning framework with manifold regularization in vector-valued Reproducing Kernel Hilbert Spaces; the similarity between two individuals is computed for each feature (descriptor) using a kernel function, and a learning algorithm is used to combine all the features, defining a mapping between their values and the identity of the individual of an input image, for a given template gallery.

2.2 Dissimilarity-based descriptors for multi-modal person re-identification

Here we summarize our MCD descriptor [83], and show how it can be extended to multi-modal re-identification.

2.2.1 Multiple Component Dissimilarity representation

Person re-identification is a *matching* problem: it basically consists of ranking a set of template images with respect to their similarity to the query image, computed as a match score. In [85] we pointed out that the descriptors used by most appearance-based re-identification methods share two high-level characteristics: they subdivide human body into parts, and represent each part as a bag (set) of low-level local features, like random patches and SIFT points. This suggested us an analogy with the kind of descriptor proposed in the Multiple Component Learning (MLCL) framework of [28] for the different task of object *detection*. It consists of representing an object (e.g., a pedestrian) as an ordered sequence of parts (or even as a single part), each of which is as an unordered set of components, and is represented by a suitable feature vector. In MCL, an object detector is constructed by combining the part detectors; each part detector is a classifier that is constructed using the Multiple Instance Learning (MIL) paradigm [27], according to the corresponding multiple component representation.

Such a high-level analogy with MCL descriptors inspired us the Multiple Component Matching (MCM) framework for constructing multiple part-multiple component descriptors for re-identification (matching) problems, which is also capable to cast the existing descriptors [85]. We then proposed the Multiple Component Dissimilarity (MCD) framework, which allows one to convert any MCM descriptor into a dissimilarity-based one [82]. The

MCD framework, which is summarised in the rest of this section, was originally proposed for speeding up the matching step of re-identification methods, enabling real-time applications [82]. Existing descriptors are indeed rather complex, and require a relatively high matching time, while MCD descriptors are simple, fixed-length vectors of real numbers, instead. We subsequently found that MCD has other advantages, among which enabling a general implementation of the novel task presented in this chapter.

The main idea underlying MCD stems from the dissimilarity representation for pattern recognition [78], originally proposed to deal with problems in which a feature vector representation is not available or is not easy to obtain, while it is possible to define a dissimilarity measure between pairs of samples (e.g., object images). A sample can thus be represented as a fixed-size vector of dissimilarity values from a predefined set of "prototype" objects. Prototypes are chosen depending on the task at hand, e.g., by clustering, by techniques derived from feature selection approaches, or even randomly [78]. In MCD we adapted the dissimilarity paradigm to multiple part-multiple component descriptors, by constructing a set of visual prototypes for each object part (e.g., the torso of an individual), which is then represented as a vector of dissimilarity values to the corresponding prototypes. The dissimilarity vectors of each part are then concatenated into an ordered dissimilarity vector representing the whole object. The main difference with the original dissimilarity representation is that, in MCD, the visual prototypes are representative of *local* characteristics of each body part, instead of the whole object. In particular, each prototype is defined as a set of components, according to the underlying MCM representation.

In [83] we exploited them to devise a framework for Multiple-Part Multiple-Component (MPMC) clothing appearance descriptors, in which (see Sect. 2.1.1): (i) A *body part subdivision* is often used, and a distinct representation is built for each part (otherwise, the whole body can be viewed as a single component); (ii) Each part (or the full body, if no part subdivision is used) is represented by *multiple components*, e.g., patches, strips, or interest points, and each component is described with a distinct feature vector. The MPMC representation can be formalized as follows. The image $\mathbf{I}$ of an individual is represented as an ordered sequence of descriptors $\{I_1,\dots,I_M\}$ of $M \geq 1$ predefined body parts. Each descriptor is a set of feature vectors extracted from n_m different components, $I_m = \{x_{m,1},\dots,x_{m,n_m}\}$, $x_{m,k} \in \mathscr{X}_m$ (different feature spaces $\mathscr{X}_m$ can also be used for different parts). An example is shown in Fig. 2.3.

Fig. **??** summarises the prototype construction procedure.

Our MCD framework was originally aimed at speeding up matching of MPMC descriptor pairs, by converting them into dissimilarity vectors. This is obtained in two steps.

Prototype selection: a "visual" prototype $\mathbf{P}_m$ is defined off-line for each body part, from a given set $\mathscr{I}$ of images of individuals. Each prototype is obtained by merging the components of the m-th body part of all images in $\mathscr{I}$ and grouping them into N_m clusters: $\mathbf{P}_m = \{P_{m,j}\}_{j=1}^{N_m}$. Each cluster is thus a set of components, $P_{m,j} = \{p_{m,j}^i\}_{i=1}^{N_{m,j}}$, $p_{m,j}^i \in \mathscr{X}_m$, and represents a specific low-level visual characteristic of the corresponding part in the feature space $\mathscr{X}_m$ (e.g., a certain color distribution).

Dissimilarity descriptor computation: from the original MPMC descriptor of any individual's image $\mathbf{I}$, a dissimilarity vector I_m^{D} is obtained for each part m as:

$$I_m^{\mathrm{D}} = \left[d(I_m, P_{m,1}),\dots,d(I_m, P_{m,N_m})\right] , \tag{2.1}$$

where the superscript D denotes a dissimilarity descriptor, and $d(\cdot,\cdot)$ is a dissimilarity mea-

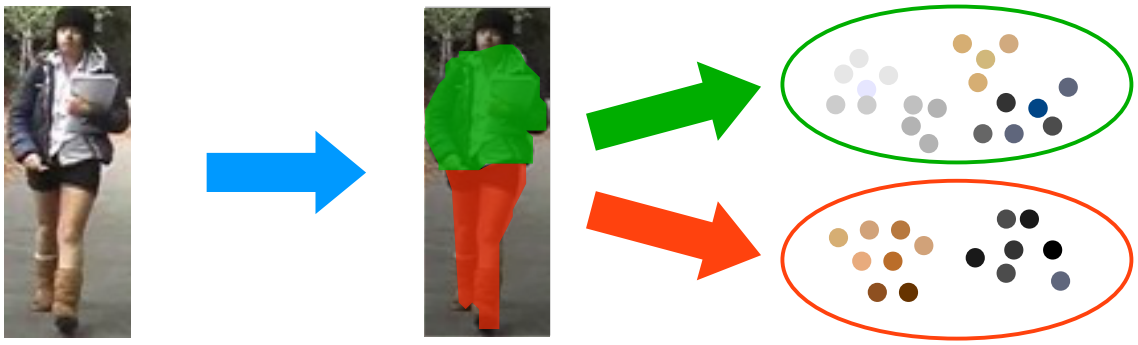

Figure 2.3: An example of MPMC representation. (a) The image of an individual. (b) The body is subdivided in two parts, upper (green) and lower body (red). (c) A set of components (e.g., image patches), sketched here as colored dots, is extracted form each part.

sure between sets of components (e.g., the Hausdorff distance [97]); the vectors I_m^{D} are then concatenated into a dissimilarity vector for the whole image:

$$\mathbf{I}^{\mathrm{D}} = [I_1^{\mathrm{D}}, \ldots, I_M^{\mathrm{D}}] \,. \tag{2.2}$$

Computing the similarity between the MCD descriptors of of a *probe* and a *template* image amounts to comparing two real-valued vectors, which can be much faster than evaluating the similarity between the original descriptors [83]. In [83] we devised a similarity measure suitable to MCD descriptors, with the following rationale: if the images of two individuals $\mathbf{I}'$ and $\mathbf{I}''$ do not exhibit the local characteristic associated with a set of components $P_{m,j}$ (i.e., both $d(I'_m, P_{m,j})$ and $d(I''_m, P_{m,j})$ exhibit high values), then $P_{m,j}$ does not provide any information about their similarity. Conversely, the smaller the values of either $d(I'_m, P_{m,j})$ or $d(I''_m, P_{m,j})$, or both, the higher the information $P_{m,j}$ conveys about the similarity between $\mathbf{I}'$ and $\mathbf{I}''$. Accordingly, we defined a weighted Euclidean distance that gives higher weights to elements that exhibit smaller values either in $\mathbf{I}'^{\mathrm{D}}$ or in $\mathbf{I}''^{\mathrm{D}}$. Assuming that $d(\cdot,\cdot) \in [0,1]$:

$$D(\mathbf{I}'^{\mathrm{D}}, \mathbf{I}''^{\mathrm{D}}) = \sqrt{\sum_{m=1}^{m} \sum_{j=1}^{N_m} \frac{w_{m,j}}{W} |d(I'_m, P_{m,j}) - d(I''_m, P_{m,j})|^2} \,, \tag{2.3}$$

where $w_{m,j} = (1 - \min\{d(I'_m, P_{m,j}), d(I''_m, P_{m,j})\})^2$, and W is a normalization factor such that $\frac{1}{W}\sum_{m=1}^{m}\sum_{j=1}^{N_m} w_{m,j} = 1$ (in [83] a different definition of w_i was used; we found that the one of Eq. (2.3) is more effective for multi-modal descriptors).

Finally, in [83] we showed that prototypes can be constructed off-line, even from a *different* set of images than the template gallery, without affecting re-identification accuracy. Furthermore, since prototype construction is an unsupervised procedure (i.e., the identity of the individual is not used), in off-line re-identification settings prototypes can be constructed using also the available *probe* images.

2.3 Semantic retrieval of pedestrian images in video surveillance scenarios

In this section we focus on a task related to person re-identification, named "semantic retrieval of pedestrians in video-surveillance scenarios". It consists of retrieving images or video sequences of individuals who match a query given in terms of a *textual* description of clothing appearance, instead of an *image*. This functionality can be very useful in applications like forensic video analysis, where the query can be obtained from the description of the suspect author of a crime made by a witness.

A similar task was considered in [96, 94, 53]. In [96, 94] it was named "person attribute search" or "attribute-based people search". The focus of [96] was on face attributes, like the presence of beard and eyeglasses, while only the dominant colour of torso and legs was considered as clothing appearance attribute. A specific detector was then developed for each attribute of interest. In [94] the following attributes were considered: gender, hair/hat colour, the position and colour of the bag (if any) carried by an individual, and, as in [96], the colour of torso and legs. A generative model was proposed to build the corresponding descriptors. In [**?**] the standard person re-identification task was considered instead, and it was argued that human experts carry it out by looking at mid-level attributes like hair style, shoe type and clothing style. Accordingly, the use of such a kind of attributes as additional features was proposed, to complement low-level features used by existing re-identification methods. The following fifteen attributes were chosen: shorts, skirt, sandals, backpack, jeans, logo, v-neck, open-outerwear, stripes, sunglasses, headphones, long-hair, short-hair, gender, carrying-object. The corresponding detectors were implemented as binary classifiers, using ad hoc features, and their output scores were fused with the match score produced by the re-identification method of [32].

Since these early works above mentioned, tremendous strides have been made in the field, towards the retrieval of more particular attributes such as clothing type, hair length and skin tone, and a better accuracy of the prediction models.

Usually, to get a semantic people description from the pedestrian image, the methods in literature start with some pre-processing technique in order to remove the background and segment the human body in parts. Background can be subtracted from videos using frame differencing, or by more sophisticated segmentation models. In particular, the background can be encapsulated into a model and subtracted to the frame containing the pedestrian, such what has been done in [101] using the technique in [57]. Adopting devices such as the Microsoft Kinect it is possible to get an accurate extraction of the pedestrians pixels from videos along with the position of the body parts [77]. Other methods separate the background from still RGB images instead of videos, using other models such as the STEL generative model in [31] or the deep decompositional network of [63] in [35] and this work. The advantage of the latter consists on a more accurate real-time background subtraction that also is able to extract different body parts from the silhouette: hairs, head, torso, legs and shoes. Not every approach needs this pre-processing step: some methods divide the body in several horizontal stripes [53] or take as input the entire color frame [56].

Once pre-processing has been done, a descriptor is extracted in order to capture low-level features and/or mid-level semantic descriptors based on soft biometrics. To describe the pedestrian appearance the most used features include color, texture, and shape. Color histograms in different color spaces are used in [84, 87] and [35] in order to extract cloth-

ing color attributes. Texture is another popular feature used to describe clothing. In [31] Recurrent High-Structured Patches (RHSPs) are used to highlight the texture characteristics that are highly recurrent in the pedestrian appearance, while in [53] texture filters (Gabor, Schmid) are extracted from the luminance channel. Other features like local key-points, shape information and group information are also used for related tasks such as person re-identification. In this work and in [56, 19] convolutional neural networks are used to extract automatically features in a supervised way. With the recent breakthrough performances of deep neural networks in vision tasks [52],this approach is one of the most promising also for attribute extraction from low resolution natural images. A convolutional neural network can be used to directly classify the attributes, but can also be used to extract features that may be useful also for related tasks.

Low and medium level features can be used to detect attributes using classifiers such as svm [87, 35, 73, 100], or deep learning algorithms [56, 19]. Some methods take also into account that certain clothing attributes are more likely to co-occur than others. For instance the gender of a person may make certain articles of clothing or types of haircut more likely than others. This has been exploited for instance by [88] employing Markov Random Field (MRF) to model these possible dependencies, and by [92] where a transformation matrix is learned, in order to convert the original binary attributes to continuous attributes. This can also take into account of situations where one or more attributes are not recognized by the classifier but can be inferred from others. For instance the attribute female can be inferred by the presence of both skirt and handbag. In this work we decided to incorporate these dependencies directly in the objective function of the deep neural network using an inference model introduced by [36].

Once attributes are extracted, they can be used not only to retrieve pedestrian and annotate images, but can also be combined to form an overall descriptor for a person in order to assist person re-identification. Attributes can be directly used to match people, such as in [101], where a Nearest Neighbour strategy is used to match pedestrian using their attributes. The attributes can also be used to refine the ranked list that compose the result of a re-identification query. For instance, in [29] this is obtained by using an adaptive similarity model based on spatially constrained attributes. Another option is to use some metric learning framework in order to produce small distances for image pairs showing the same person and large distances for image pairs showing different persons. For instance, in [88] a Logistic Discriminant Metric Learning (LDML) is applied to automatically determine a distance metric for all attributes.

Chapter 3

Multimodal Person Re-Identification Using RGB-D Cameras

Person re-identification consists of recognizing individuals across different sensors of a camera network. Whereas clothing appearance cues are widely used, other modalities could be exploited as additional information sources, like anthropometric measures and gait. In this chapter we focus on anthropometric measures, since their extraction has been eased by recently introduced RGB-D sensors. This makes it possible to work in unconstrained re-identification settings, whereas RGB sensors require complex calibration procedures, and are very sensible to occlusions, clutter and lighting conditions.

We address in particular two issues, that have not been considered in depth in previous work. (1) Can the re-identification accuracy of clothing appearance be improved by fusing it with anthropometric cues, in *unconstrained* re-identification settings? We address this issue by selecting anthropometric measures that can be extracted in such settings, using commercial RGB-D sensors like the Kinect, among the ones proposed in previous work; we then fuse them with different state-of-the-art clothing appearance descriptors. (2) How to combine descriptors coming from different modalities? All previous works used *score-level* fusion rules. Here we also explore a different possibility, by developing a fusion method based on *feature-level* fusion, extending our previously proposed Multiple Component Dissimilarity (MCD) descriptor (see section 2.2). We originally developed MCD for reducing matching time of clothing appearance descriptors; here we show that it exhibits some interesting properties for multi-modal fusion.

The experimental evaluation is carried out on two data sets including RGB-D data, one of which is a novel, publicly available data set that we acquired using Kinect sensors. The fusion with anthropometric measures increases the first-rank recognition rate of clothing appearance descriptors up to 20%, whereas our fusion approach reduces the processing cost of the matching phase.

The remainder of the chapter is structured as follows. In Sect. 3.1 we describe whe extension of our dissimilarity-based framework for multi-modal descriptors. In Sect. 3.2 we choose a set of anthropometric measures that can be extracted from RGB-D data in unconstrained settings, and describe the clothing appearance descriptors used in the experiments. Experimental results are reported in Sect. 3.3. Conclusions and suggestions for future research directions are given in Sect. 3.4.

3.1 Extending MCD descriptors to multi-modal person re-identification

Here we show that our MCD framework can be applied to descriptors of other modalities beside clothing appearance, and that it provides an interesting solution for fusing descriptors of different modalities in a feature-level fashion. To this end, a given descriptor has first to be framed as a multiple-component multiple-part (MPMC) one. In particular, a given set of q anthropometric measures can be seen as the simplest MPMC representation made up of one part (the whole body, $M = 1$) and one component, the feature vector $x_{1,1} \in \mathscr{X}_1 \subset \mathbb{R}^q$, so that the corresponding descriptor of an individual $\mathbf{I}$ is given by $I_1 = \{x_{1,1}\}$. According to section 2.2, to obtain an MCD descriptor one has to construct one set of prototypes $\mathbf{P}_1$, which is made up of a set of N_1 clusters, $\mathbf{P}_1 = \{P_{1,j}\}_{j=1}^{N_1}$, obtained by grouping the vectors of anthropometric measures of a given set of individuals $\mathscr{I}$; each cluster is a set of components $P_{1,j} = \{p_{1,j}^i\}_{i=1}^{N_{1,j}}$, where each $p_{1,j}^i \in \mathscr{X}_1$ is the vector of anthropometric measures of a given individual in $\mathscr{I}$. The corresponding MCD descriptor will be given by

$$\mathbf{I}^{\mathrm{D}} = [I_1^{\mathrm{D}}] = \left[d(I_1, P_{1,1}), \ldots, d(I_1, P_{1,N_1})\right]. \tag{3.1}$$

In general, given $K > 1$ different modalities and their original descriptors, one obtains K dissimilarity vectors, each defined as in Eq. (2.2).

Let's consider now the issue of how to combine the descriptors of K different modalities. In re-identification tasks one can use either feature-level fusion (concatenating the feature vectors of each modality into a single one, and then computing an overall matching score), or score-level fusion (fusing the matching scores computed separately for each modality). As pointed out in Sect. 2.1.3, in all previous works on multi-modal re-identification score-level fusion was used, although the fusion method of [33] can also be used for this purpose. Score-level fusion appears as the most straightforward solution, for at least two reasons. One reason is that in multi-modal systems feature-level fusion requires one to concatenate heterogeneous quantities, like a color histogram for clothing appearance and a set of anthropometric measures (the same issue arises, e.g., in multi-modal biometric identity recognition [81]), whereas score-level fusion allows one to combine homogeneous information, i.e., a set of similarity scores (one for each modality). Another reason is that descriptors that lie in high-dimensional feature spaces (like many clothing appearance ones) may overwhelm the contribution of descriptors that lie in relatively lower dimensional spaces (e.g., a vector made up of a few anthropometric measures). Our dissimilarity-based MCD descriptor provides however a different perspective for feature-level fusion. First, dissimilarity-based descriptors are *representation-independent*, i.e., they are logically and semantically at a higher level than the underlying object representation. This implies that dissimilarity values computed on different modalities are semantically as coherent as the matching scores computed from the original descriptors: in the case of MCD, the only difference is that they encode a (dis)similarity between *local* object components (e.g., body parts) instead of between *whole* objects. Second, the size of MCD descriptors can be controlled by setting the desired number of prototypes (see Sect. 2.2.1), independently of the feature set size of the original descriptors: this allows one to avoid concatenating vectors of very different size. Obviously, reducing the number of prototypes below a certain amount may affect the resulting re-identification accuracy: this issue will be empirically investigated in Sect. 3.3. To sum up, feature-level fu-

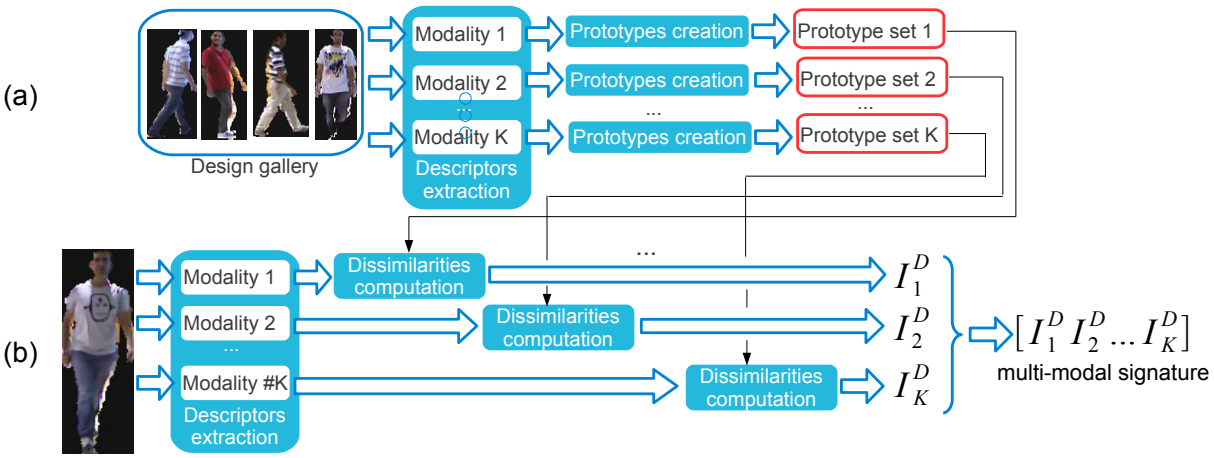

Figure 3.1: Outline of our multi-modal MCD representation. (a) *Prototype construction*: a MCMP descriptor for each modality is extracted from a design set of individuals' images, and a distinct set of prototypes for each modality is constructed. (b) *Multi-modal descriptor computation*: an image of an individual is represented in the dissimilarity spaces associated with each set of prototypes, and the resulting dissimilarity vectors are concatenated.

sion of MCD descriptors is in principle not affected by the issues that affect non-dissimilarity descriptors.

Our MCD framework can be extended to K different modalities as follows. A distinct prototype set is first constructed for each of modality. Then, given the image of an individual and the K original descriptors, the corresponding dissimilarity vectors $I^{\mathrm{D},k}$, $k = 1,\ldots,K$, are computed, and the final MCD descriptor is obtained by concatenating them in any predefined order, e.g.:

$$\mathbf{I}_{\mathrm{m}}^{\mathrm{D}} = [\mathbf{I}^{\mathrm{D},1},\ldots,\mathbf{I}^{\mathrm{D},K}] \,. \tag{3.2}$$

The matching score between two descriptors can finally be computed using again Eq. (2.3). The proposed multi-modal MCD representation is summarized in Fig. 3.1.

3.2 Clothing appearance descriptors and anthropometric measures

To investigate the two issues mentioned in the beginning of the chapter, here we explain our choice of a set of anthropometric measures that can be extracted from RGB-D sensors from unconstrained poses, and of clothing appearance descriptors. We also describe the construction of the corresponding MCD descriptors.

3.2.1 Anthropometric measures

The depth map and the person detection functionality provided by off-the-shelf RGB-D sensors enable a relatively easy detection of some anthropometric measures (see Sect. 2.1.2). Further ones can be extracted exploiting the pose estimation functionality of the Kinect SDK (based on converting depth data to cloud points in real-world coordinates), which provides a real-time estimation of the absolute position in metric coordinates of 20 different body joints

(see Fig. 2.2): the spine, the centre of the hip, shoulder and head, left and right shoulder, elbow, wrist, hand, hip, knee, ankle and foot. This allows further measures to be extracted. Each joint is associated with a tracking state: "tracked", if it is clearly visible, and "inferred" if it is not, but the Kinect SDK can infer its position. The joints that can be actually tracked or inferred, and the precision of their localization, depend on the pose; e.g., in a lateral pose the joint of the farthest shoulder is usually not tracked.

To the purpose of this work, we selected a set of anthropometric measures among the ones proposed in previous work (see Sect. 2.1.2), focusing on measures that can be extracted from *unconstrained* poses, and with low processing cost, to fit real-world video surveillance and re-identification scenarios. For instance, this is the case of the height of a person: it can be extracted from the silhouette, e.g., by measuring the distance between the highest silhouette point and the floor plane, in real-word coordinates. This is not the case of geodesic distances of [7]: they were estimated from the 3D mesh of the abdomen, which can be extracted only from a frontal pose, and with a relatively higher complexity. Specific issues also arise for measures that can be extracted from skeleton joint positions. The positions of some joints are estimated more reliably from frontal poses (probably because the Kinect device has been designed for tracking individuals standing in front of the sensor), and may change significantly across different poses; e.g., in a rear pose the head and the neck joint are usually localized in an higher position than in a frontal pose (see Fig. 2.2(d)). Moreover, although the distances between all pairs of adjacent joints could be used as anthropometric measures, some pairs of joints are closer than others (e.g., the hip width and the shoulder width), and thus their estimated distance may be affected by a higher relative error and exhibit a higher variance. Accordingly, we selected the following seven measures from [7] and [69] (denoted as d_1 to d_7), and two measures (d_8 and d_9) from [38]:

- d_1 distance between floor and head
- d_2 ratio between torso and legs
- d_3 height (distance between the highest body silhouette point and the floor plane)
- d_4 distance between floor and neck
- d_5 distance between neck and shoulder
- d_6 distance between torso centre and shoulder
- d_7 distance between torso centre and hip
- d_8 arm length (sum of the distances between shoulder and elbow, and between the elbow and wrist)
- d_9 leg length (sum of the distances between hip and knee, and between knee and ankle)

All distances are Euclidean. In particular, d_6 and d_7 were computed as geodesic distances in [7], but we replaced them with Euclidean ones to make them pose-invariant; we also averaged the pairs of measures exhibiting vertical symmetry (d_5, d_6 and d_7), if both are in the "tracked" status; otherwise we used only the "tracked" one. Note also that we computed d_2 as in [7]: $d_2 = \frac{d_5}{d_1 \cdot d_{\text{floor-hip}}}$. We finally normalized all these measures (both in template and in

probe images) to zero mean and unit variance (mean and variance were computed on the template gallery).

According to Sect. 3.1, we built a MPMC descriptor of anthropometric measures made up of one body part and one component. The latter is represented as a vector $x = [d_1, \ldots, d_9]$. In [7] each value was computed from the video frame exhibiting the highest number of joints with status "tracked". To improve robustness, we computed each value as the median over the first ten frames from a video. We will compare the two strategies in Sect. 3.3.2. We then computed the matching score s between the descriptors of two individuals x' and x'' as in [7], using a weighted Euclidean distance to take into account the different discriminant capability of each measure:

$$s = \sum_k w_k (d'_k - d''_k)^2 \,, \tag{3.3}$$

with $w_k \geq 0$ and $\sum_k w_k = 1$. Details about weight computation are given in Sect. 3.3.2.

3.2.2 Clothing appearance descriptors

We chose three MPMC clothing appearance descriptors: the state-of-the-art SDALF [31] and eBiCov [64], and our MCMimpl [85], which we used in our first work on the MCD framework [82]. We implemented SDALF and eBiCov using the source code provided by the authors. SDALF subdivides the body into torso and legs through a horizontal axis that is found by exploiting symmetry and anti-symmetry properties of the silhouette's color and shape, and used three kinds of features. *Maximally Stable Color Regions* (MSCR) are non-regular regions of homogeneous color, extracted from the whole body, which describe the per-region color displacement, and are found via agglomerative clustering; each one is represented by its area, centroid, second moment matrix and average color, resulting in a 9-dimensional vector. *Recurrent High-Structured Patches* (RHSP) are rectangular patches made up of recurrent, repeated patterns, separately extracted from each part, and represented by a rotation-invariant LBP histogram; they highlight texture characteristics that are highly recurrent in the pedestrian appearance. Both MSCR and RHSP are sampled mainly around the vertical axis of symmetry of each body part. Weighted HSV histograms (w-HSV) are extracted from each body part to capture the chromatic content (giving lower weights to pixels closer to the body periphery), and are concatenated into a single feature vector. SDALF can be conveniently seen as being made up of $M = 4$ sets of components: the MSCR feature vector, the RHSP feature vectors extracted from torso and legs, and the concatenated HSV color histogram.

The eBiCov ("enriched gBiCov") descriptor [64] combines SDALF with the gBiCov descriptor. gBiCov is made up of a Biologically Inspired Features (BI) [79] and a Covariance (COV) descriptor. Two layers were selected from BI: Gabor filters and the MAX operator, for improving respectively the robustness to illumination changes, and to scale changes and image shifts. COV is used to compute the similarity of BI features taken at neighboring scales, capturing shape, location and color information. Each of them is extracted from the whole body (without background subtraction) separately from the three HSV channels, and the three resulting feature vectors are then concatenated. Therefore, gBiCov can be seen as made up of two components (BI and COV) extracted from a single part, and eBiCov as a descriptor made up of $M = 6$ components (4 for SDALF and 2 for gBiCov).

The original MCMimpl descriptor uses the same body subdivision as SDALF. In this work we used an enhanced version, exploiting the skeleton points extracted by the Kinect SDK:

using only points that can be detected from *any* pose (see Fig. 2.2(b),(c),(d)), we subdivided the body into $M = 4$ parts: upper and lower torso, upper and lower legs. The torso region is localized as the portion of the image between the y coordinates of shoulder and hip centers. The mask pixels corresponding to the first half of the torso region are considered as the upper torso, and the other ones as the lower torso. The mask pixels between the coordinate of the hip centre and the average of the y coordinates of the knees (or the y coordinate of the visible knee, if only one is detected), define the upper legs region. The mask pixels between the average y coordinate of the knees and the bottom of the mask define the lower leg region. The set of components of each body part is obtained by randomly extracting image patches of different sizes, possibly overlapping; each patch is represented with an HSV color histogram.

We refer the reader to [31, 85, 64] for further details.

3.2.3 Computing MCD descriptors

To obtain a MCD descriptor we had to choose a clustering technique for prototype construction, and a distance measure $d(\cdot,\cdot)$ between sets of components (see Eq. 2.1). For non-singleton sets of components (e.g., the HSV histogram in MCMimpl), we used the two-stage clustering approach of [83], aimed at reducing the processing cost. It consists of a first Mean-Shift clustering step [21] applied to each set of components, and a subsequent c-Means step applied to the first-stage centroids. For singleton set of components (e.g., the vector of anthropometric measures), only the c-Means step was carried out. The prototypes were defined as the resulting c centroids. In our experiments we set the bandwidth parameter of Mean-Shift to 0.3, $c = 200$ for clothing appearance descriptors, and $c = 30$ for anthropometric measures. This choice of c is discussed in Sect. 3.3, where we evaluate how the number of prototypes affects the re-identification accuracy.

We defined $d(\cdot,\cdot)$ as the modified k-th Hausdorff distance [97], which is known to be robust to outliers. It is defined as the k-th ranked minimum distance between all pairs of elements from two sets P and Q:

$$d(P,Q) = \max\{h_k(P,Q), h_k(P,Q)\}\,, \tag{3.4}$$

where

$$h_k(P,Q) = k\text{-th} \min_{p\in P, q\in Q} (\|p-q\|)\,. \tag{3.5}$$

The parameter k controls the influence of outliers. We set $k = 10$. We then chose the same distance metric $\|\cdot\|$ between components, both for the SDALF and MCMimpl descriptors (we refer the reader to [31, 85] for further details). For the chosen anthropometric descriptor, the set of components is a singleton (a single feature vector). Hence, we defined $\|\cdot\|$ as the weighted distance

$$\|x' - x''\| = \sum_k w_k (d'_k - d''_k)^2\,, \tag{3.6}$$

using the same weights w_k of Eq. (3.3) to take into account the discriminant capability of the different measures.

3.3 Experimental evaluation

According to our choice of anthropometry measures and clothing appearance descriptors, the goals of our experiments are the following: (1) Evaluating whether anthropometric cues can improve the re-identification accuracy of clothing appearance ones in unconstrained re-identification settings; to this end, we carried out experiments on the three clothing appearance descriptors mentioned above. (2) Evaluating two different techniques for combining multi-modal descriptors: score-level fusion, and our dissimilarity-based MCD feature-level fusion. We describe in Sect. 3.3.1 the data sets we used and the experimental setup, in Sect. 3.3.2 the weights assigned to each of the chosen anthropometric measure, and in Sect. 3.3.3 the experimental results.

3.3.1 Data set and experimental setup

To carry out our experiments, a dataset including RGB and depth data (the estimated positions of the joints) is required. Most benchmark data sets for person re-identification were acquired using only RGB sensors, e.g., [42, 48, 89]. To our knowledge, the only data set that contains also RGB-D data and can be used for our purposes is "RGBD-ID" [7]; since it was designed mainly for using depth data (sometimes the same individual wears different clothes in different acquisitions), we modified it as described below. The BIWI RGBD-ID data set of [69] was designed for a long-term setting, and therefore most of the individuals wear different clothes in training and testing sequences; it is thus not suited to short-term re-identification settings including clothing appearance descriptors. Beside using the data set of [7], we also acquired a new data set of video sequences, named "KinectREID", which is available upon request.[1]

Our KinectREID data set was acquired using Kinect sensors and the official Microsoft SDK. It consists of video sequences of 71 individuals taken at a lecture hall of our department, under different lighting conditions and three view points: three near-frontal views, three near-rear views, and one lateral view. All the individuals were requested to walk normally along a predefined path; some of them carried accessories like bags. Seven video sequences were taken for each individual, for a total of 483 video sequences (14 sequences showing lateral poses, on which the SDK tracking failed, were discarded). Each sequence lasts for about 10 sec., but depth data is available only for a few seconds, corresponding to the range of the Kinect device (about 0.8 to 4.0 m). Some examples are shown in Fig. 3.2. Each tracked individual was associated with a sequence of frames, with the corresponding segmentation masks and skeleton points. We used both the RGB frames and the skeleton points to extract one clothing appearance descriptor from each frame, and the skeleton points to estimate the anthropometric measures.

RGBD-ID [7] was acquired using Kinect cameras as well, but with the OpenNI SDK.[2] It contains RGB and depth data for 80 individuals. Differently from KinectREID, in RGBD-ID: (i) the joint positions were obtained using another tracker, and wrist and ankle are not included; (ii) four acquisitions were made for each individual, one rear and three frontal poses, and in one of the latter the arms are stretched; (iii) for each acquisition only 4 or 5 RGB-D frames are provided; (iv) sometimes, the same individual wears different clothes in different acquisitions: we removed the corresponding tracks. Hence, 2 to 4 acquisitions

[1] More information at http://pralab.diee.unica.it/en/PersonReIdentification
[2] http://structure.io/openni

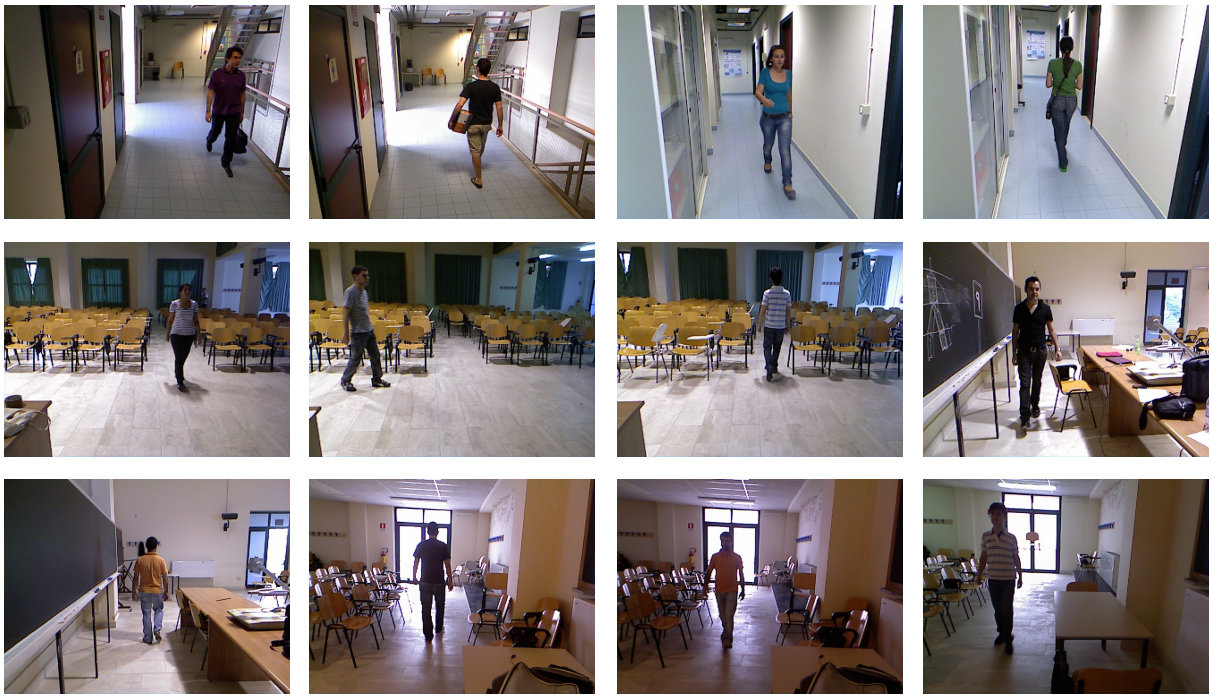

Figure 3.2: Frames taken from our KinectREID data set. Note the different view points (camera angle and position), locations, poses and illumination.

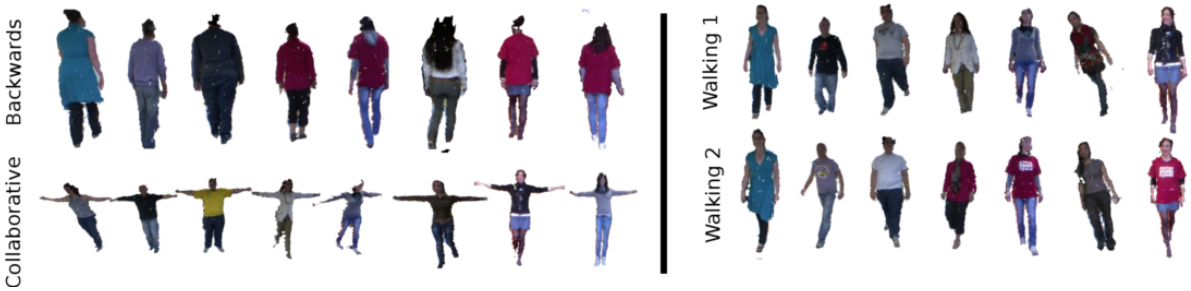

Figure 3.3: Frames taken from the RGBD-ID data set. Note the different poses, and the presence of different subjects wearing the same t-shirt.

remained for each individual, for a total of 197 video sequences out of the 320 original ones. Some examples are shown in figure 3.3

We carried out our experiments simulating a *closed-set* scenario, as in most of the existing works on person re-identification. In this scenario, the identity of each probe individual corresponds to one of the template identities. On both data sets, the experimental setup was the following. First, we selected 20 individuals for estimating the weights of the anthropometric measures (see Sect. 3.3.2). These individuals were not used in the subsequent steps. For each of the remaining individuals, one video sequence was randomly chosen as a template track, and the remaining ones were used as probes. MCD prototypes were computed on the template gallery: $c = 200$ prototypes were used for clothing appearance descriptors, and $c = 30$ for anthropometric measures (see Sect. 3.3.3). For matching a pair of probe and template tracks we used the *multiple shots vs. multiple shots* (MvsM) setup [31], i.e., we matched several pairs of the respective frames. To reduce processing cost we discarded from KinectREID the frames in which the skeleton was not available, and chose the first 10 remaining

frames from each track. We used all the available frames (4 or 5) from RGBD-ID, instead. We evaluated the matching score between each pair of frames using Eq. (3.4) for the original descriptors, and Eq. (2.3) for MCD descriptors, and used their median value as the final matching score. The score-level fusion of the original clothing appearance descriptors was implemented as the widely used sum rule, which in preliminary experiments attained better results than other rules (minimum, maximum and product of the scores). We repeated the above procedure for ten times, and evaluated the results as the average Cumulative Matching Characteristics (CMC) curve: it is defined as the probability of finding the correct match within the first n ranks, with n ranging from 1 to the number of templates. All the frames used as probes and templates in each run, and the corresponding skeletal points, are available upon request (see the URL above).

3.3.2 Combination of the anthropometric measures

As explained in Sect. 3.2.1, to compute a matching score between two anthropometric descriptors (in our case, two vectors of anthropometric measures), and their similarity measure in the case of the corresponding MCD descriptor, we used a weighted combination of the normalized anthropometric measures, respectively Eq. (3.3) and (3.6). We computed the weights separately on the two data sets. To this end we maximized the $AUC_{20\%}$ performance index (the area of the first 20% ranks of the CMC curve, normalized to $[0,1]$) obtained by computing the matching score as in Eq. (3.3), on the subset of 20 individuals mentioned above. To find the "optimal" weights we used the *quasi-exhaustive* strategy of [7], i.e., a grid search in the weight space, considering for each weight the values from 0 to 1 with step 0.05. We also compared the two strategies mentioned in Sect. 3.2.1 for computing the matching score: using only the pair of probe and template frames exhibiting the highest number of joints with status "tracked" as in [7], and computing the median over all the considered frames (10 on KinectREID, 4 or 5 on RGBD-ID).

In both data sets we obtained the highest $AUC_{20\%}$ value by computing the median score among all the available frames: 57% vs 43% on KinectREID, and 60% vs 57% on RGBD-ID (the lower improvement on RGBD-ID is probably due to the lower number of available frames). Accordingly, we used this strategy in the rest of the experiments. The weight values (see Sect. 3.2.1 for the corresponding anthropometric measures) were the following. On KinectREID: $w_2 = 0.2$, $w_3 = 0.5$, $w_4 = 0.05$, $w_8 = 0.05$, $w_9 = 0.2$, whereas $w_1 = w_5 = w_6 = w_7 = 0$. Probably the distance between floor and head (d_1) was not discriminant, since the head joint in the rear pose is usually tracked in an higher position by the Kinect SDK, whereas the distances between near joints (neck to shoulders d_5, torso centre to shoulders d_6, and torso centre to hips d_7) are probably too small and therefore subject to higher relative errors. Some measures may also be redundant, due to a high correlation with other ones. On RGBD-ID we obtained the following weights: $w_1 = 0.4$, $w_3 = 0.6$, $w_8 = 0.05$ and zero for the remaining ones. The differences between the weights computed on the two data sets are due to the different acquisition settings, as well as on the different SDK (e.g., probably the distance d_1 between floor and head is computed differently in the Microsoft and OpenNI SDKs).

3.3.3 Experimental results

We first evaluate the discriminant capability of the anthropometric measures, both individually and jointly. For the k-th measure alone, the matching score was computed as $(d'_k - d''_k)^2$.

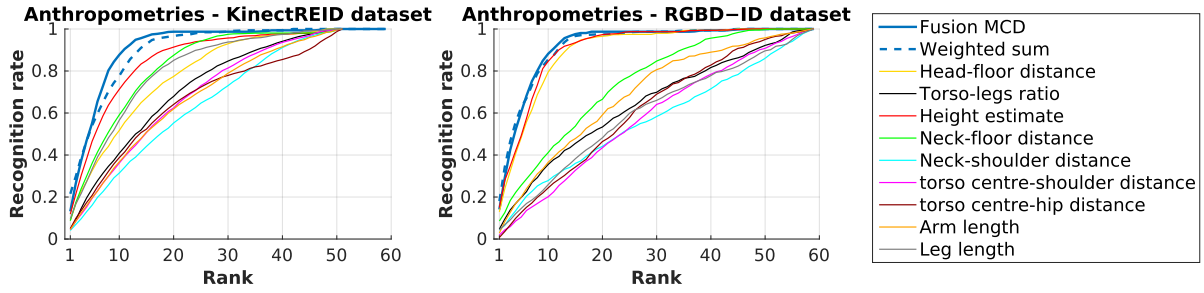

Figure 3.4: CMC curves of the individual anthropometric measures d_1–d_9 (see Sect. 3.2.1 for the definition of each measure) and of their combination (blue lines: dashed: original descriptor, solid: MCD descriptor) on the KinectREID (left) and RGBD-ID (right) data sets.

For all measures, we computed the matching score using the weights reported in Sect. 3.3.2, both for the original MPMC descriptor and for its MCD version. The CMC curves for both data sets are reported in Fig. 3.4. As one could expect, the individual anthropometric measures exhibit a very different range of performance. In particular, the height estimated as the distance between the highest body silhouette point and the floor plane (d_3) exhibited a very good performance on both data sets, if compared to the one of clothing appearance descriptors (see below, and Fig. 3.5). The combination of all anthropometric measures (the ones with non-zero weight) attained on both data sets a better performance than each individual one; sometimes the performance was similar or even better than the one of clothing appearance descriptors. Note also that the original descriptor and the MCD one exhibited a similar performance on RGBD-ID, and that the latter outperformed the former on KinectREID for ranks from about 5 to 20.

The CMC curves of the clothing appearance descriptors and of their fusion with the anthropometric ones are reported in Fig. 3.5 (the plots in each row correspond to one of the clothing appearance descriptors, the plots in each column to one data set). We first compare the CMC of each of the three clothing appearance descriptors (either the original or the MCD one) with the corresponding multi-modal fusion with the anthropometric descriptor (i.e., each pair of lines of identical color in each plot of Fig. 3.5), which corresponds to the objective (1) mentioned at the beginning of this section. It can be seen that fusing the two modalities always produced a remarkable performance improvement over the clothing appearance descriptor alone, across a large range of ranks. The only exception is the MCMimpl descriptor on RGBD-ID, where the performance improvement is limited to the first few ranks, which are nevertheless the most relevant ones for re-identification tasks. This provides evidence that anthropometric measures actually provide complementary discriminant information with respect to clothing appearance cues. In particular, we observed that they allow discriminating between different template individuals wearing clothes similar to the probe: this is the reason of the improvement in recognition rate observed in all our experiments, already at rank 1.

We now compare the score-level fusion of the original descriptors with our feature-level MCD fusion technique (i.e., the pair of solid lines in each plot), which corresponds to our objective (2). Both fusion techniques attained a similar performance; the only exception is again MCMimpl on RGBD-ID, where the score-level fusion outperformed the MCD feature-level fusion for the lowest ranks. However, we point out that the MCD fusion technique has the advantage of a much lower processing cost for the matching phase of the clothing

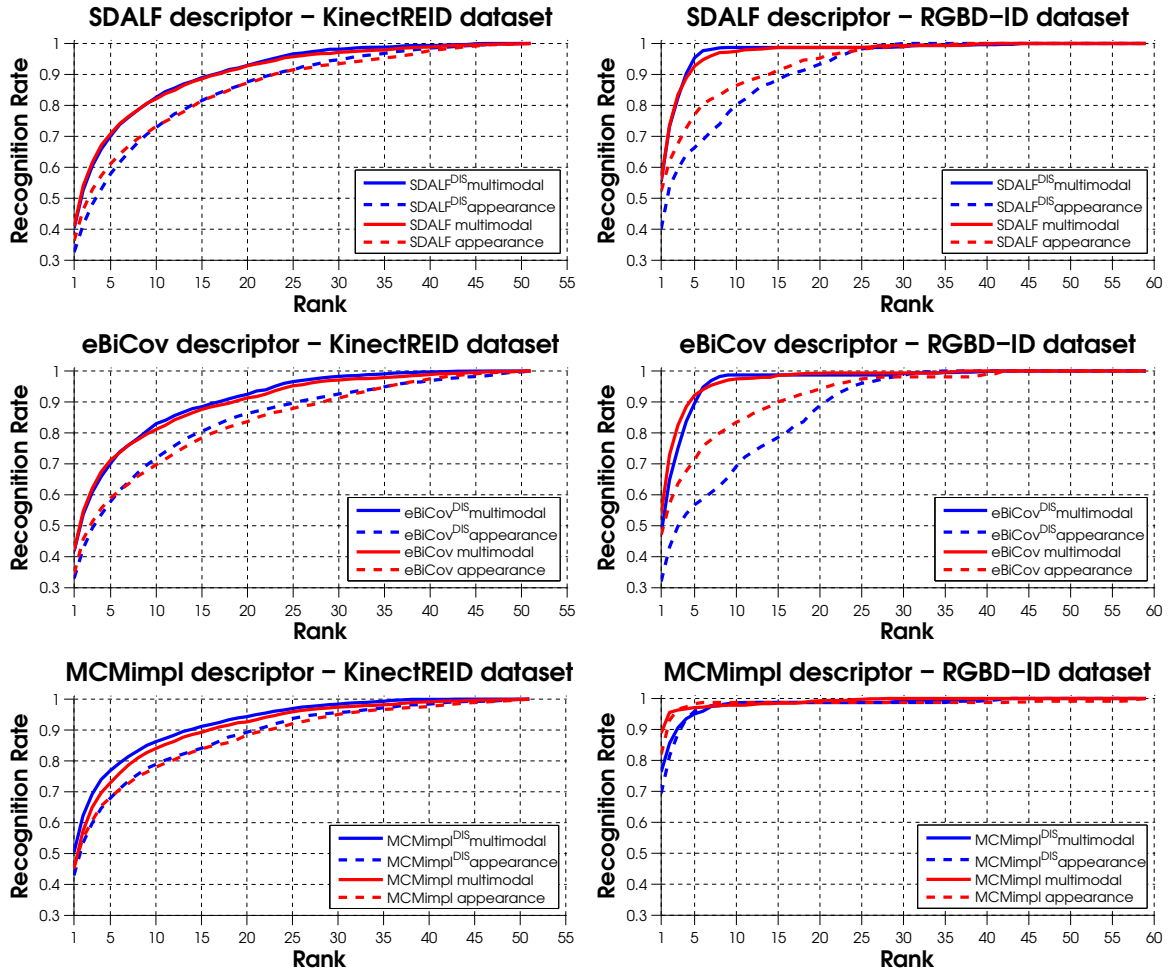

Figure 3.5: CMC curves (left: KinectREID data set; right: RGBD-ID data set) attained by the three clothing appearance descriptors (from top to bottom: SDALF [31], eBiCov [64], MCMimpl [85]), in their original (dashed red lines) and MCD [83] version (dashed blue lines), and by their fusion with anthropometric descriptors (original descriptors: solid red lines; MCD descriptors: solid blue lines). MCD descriptors are denoted with the superscript "DIS".

appearance component of the dissimilarity vector, as explained in Sect. 2.2.1. This makes our MCD fusion technique suitable for real-time multi-modal person re-identification; e.g., on a laptop with a dual core i5-2410M processor, computing one descriptor from a single frame (including all preprocessing steps) took about 50 msec., whereas matching one probe with one template track took about 0.03 msec.

For the sake of completeness, we also compare the original clothing appearance descriptors and the corresponding MCD ones (i.e., the two dashed lines in each plot of Fig. 3.5). MCD descriptors of clothing appearance exhibited a similar (e.g., for SDALF and MCMimpl on KinectREID) or lower performance (SDALF and eBiCov on RGBD-ID) than the original ones. This is in agreement with our previous results [82]: in the case of clothing appearance clues, MCD descriptors can be useful to attain a trade-off between re-identification accuracy and processing cost; sometimes, they improve both.

We finally discuss how the size of prototypes affects the accuracy of MCD descriptors. The number of prototypes depends on the value of the c parameter of the c-Means cluster-

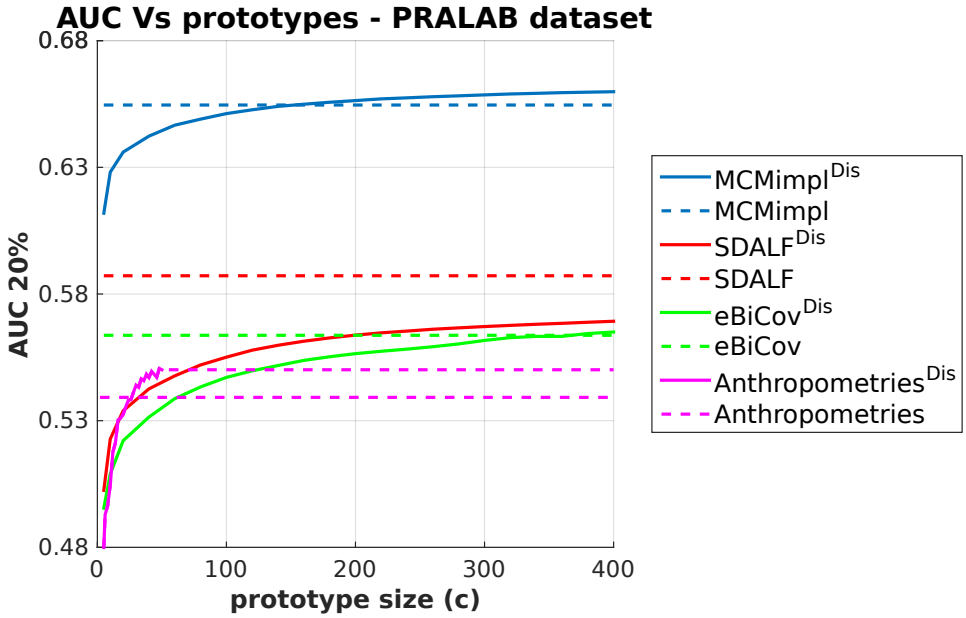

Figure 3.6: Normalized $AUC_{20\%}$ as a function of prototype size c, attained on KinectREID by the MCD clothing appearance and anthropometric descriptors (solid lines). For reference, the $AUC_{20\%}$ of the original descriptors is also shown (dashed lines).

ing algorithm we used for prototype construction (see Sect. 3.2.3); obviously, their number also affects the processing time for dissimilarity computation [83]. To get a concise overview, we report in Fig. 3.6 the average $AUC_{20\%}$ attained by MCD clothing appearance and anthropometric descriptors (solid lines) on KinectREID, as a function of c. Similar results were observed on RGBD-ID. Note that the $AUC_{20\%}$ of the anthropometric descriptor ends at $c = 51$: the reason is that no more than 51 prototypes can be obtained in MCD, since this MCD descriptor is made up of a single component and the number of template individuals is 51. For reference, the $AUC_{20\%}$ values of the original descriptors (which do not depend on c) are also reported (dashed lines). For both modalities, the $AUC_{20\%}$ initially grows as the number of prototypes increases, and attains a nearly constant value beyond a certain value of c. This value is about 200 for all clothing appearance descriptors, and about 30 for the anthropometric descriptor. These are the values of c that we used in our experiments. These results suggest that a relatively small number of prototypes can provide a good trade-off between re-identification accuracy and processing time in our MCD descriptors.

3.4 Conclusions

We investigated whether anthropometric measures can improve the re-identification performance of the widely used clothing appearance cue, in unconstrained settings, exploiting the depth information and the related functionality (in particular, the estimation of joint positions) provided by recently introduced RGB-D sensors. To this end we chose a subset of anthropometric measures proposed by other authors, which can be computed from unconstrained poses, and considered three different clothing appearance descriptors. The multimodal fusion of the two cues always attained a better performance than the clothing appearance cue alone, providing evidence that anthropometric measures provide complemen-

tary discriminant information; in particular, they allow discriminating between template individuals wearing clothes similar to the probe. We also proposed a novel dissimilarity-based, feature-level fusion technique for multi-modal re-identification, based on our MCD descriptor previously proposed for clothing appearance, as an alternative to score-level fusion, which is the only technique used so far for multi-modal re-identification. We showed that our technique can attain a better trade-off between re-identification accuracy and processing cost, when complex descriptors are involved (like clothing appearance ones). As a by-product, we acquired a novel, publicly available data set of video sequences with Kinect sensors, including both RGB and depth data.

Several future research directions can be envisaged, in the context of multi-modal re-identification using RGB-D cameras, e.g.: (i) Investigating a wider range of anthropometric cues to further improve re-identification accuracy; (ii) Developing a framework that takes into account missing cues or modalities, due, e.g., to occlusions or to the pose of an individual; (iii) Experimentally comparing the fusion technique of [33] and our MCD-based technique; (iv) Investigating the use of other modalities beside clothing appearance and anthropometric measures, as well as the fusion of different descriptors of clothing appearance (which has been already addressed in [33]); e.g., skeleton-based gait [45] could be an effective cue, whose extraction is enabled as well by RGB-D sensors, whereas remote face recognition could provide some useful cues in the case when of both the template and the probe are in frontal pose [74].

Chapter 4

Semantic retrieval of pedestrians in video surveillance scenarios

Person re-identification consists of searching for an individual of interest in video sequences acquired by a camera network, using an *image* of that individual as a query. Here we consider a related task, named *semantic retrieval of pedestrians in video surveillance scenarios*, which consists of searching images of individuals using a *textual description* of clothing appearance as a query, given by a Boolean combination of predefined attributes. This task is interesting because of applications such as forensic video analysis, where the query can be obtained from a eyewitness report.

In this chapter we develop a general method for implementing semantic retrieval of pedestrians as an extension of a given person re-identification system, using the *same* clothing appearance descriptor. We exploit the fact that most of the existing re-identification methods are based on appearance descriptors that use multiple components, and possibly a body part subdivision. Our Multiple Component Dissimilarity (MCD) framework [82] can then be used to convert any descriptor of this kind into a dissimilarity-based one, i.e., a fixed-length vector made up of dissimilarity values between a given body part and a set of visual prototypes that encode specific characteristics of that part. This allows us to devise a simple implementation, by first identifying a set of basic attributes related to visual clothing appearance, that can be detected by the original descriptor at hand (e.g., the color and the texture of upper and lower garments, the presence of short or long sleeves, etc.), and then building a detector for each attribute, as a binary classifier whose input features are made up by a dissimilarity vector. Contrary to previous work such as [96, 94, 53], we focus only on clothing appearance, and do not consider a predefined set of attributes nor a specific feature set for implementing detectors. In our method, each attribute corresponds to what we call a textual "basic query". We also propose a method for processing complex queries, obtained by combining basic ones through Boolean operators. Finally, we experimentally evaluate our method on a benchmark data set originally built for re-identification tasks.

We describe our approach for implementing semantic retrieval of pedestrians in Sect. 4.1, and experimentally evaluate it in Sect. 4.2 on a hand labelled benchmark data set for person re-identification. Sect. 4.3 concludes the chapter with some insights on future work.

4.1 A general method for retrieving pedestrians by semantic queries

Here we present a simple and general approach to implement semantic retrieval of pedestrians using textual queries related to clothing appearance attributes, based on MCD descriptors. Our approach is intended as an extension of any given re-identification method that uses images as queries, and uses a multiple part-multiple component representation of clothing appearance. It can be summarised as follows:

1. Given any multiple part-multiple component descriptor, the first step consists of identifying a set of "basic" attributes it can detect, related to visual clothing characteristics.

2. A detector is built for each attribute, using the MCD version of the descriptor at hand as input.

3. Each attribute is associated to what we call a *basic query*, i.e., an "atomic" textual query related to the corresponding visual characteristic. Complex queries can then be constructed by combining basic ones with Boolean operators, and have to be processed by suitably fusing the outputs of individual detectors.

In the following we discuss the three points above.

Identifying the set of attributes

The attributes have to be defined on the basis of basic characteristics of clothing appearance that the descriptor at hand should be capable to detect. To this aim, each predefined body part (if more than one) can also be considered separately. For instance, if a descriptor subdivides the body into upper (torso and arms) and lower (legs) parts, and uses only colour features (e.g., the HSV colour histogram), some of the attributes can refer to the colour of torso/arms, and other attributes to the colour of legs, e.g., "red trousers/skirt", and "blue upper body garment". Such a descriptor could also enable the detection of attributes like short sleeves and shorts, through skin colour. More refined body part subdivisions can enable the definition of less coarse attributes. For instance, if the descriptor separates the arms from the torso, distinct attributes related to their colours can be considered. This may allow one to retrieve, e.g., images of individuals wearing a black jacket over a white shirt, by combining the corresponding basic queries (see below). Note that the definition of attributes cannot be an automated process, but should involve human judgement. It is indeed necessary to identify which characteristics can be reasonably detected using the descriptor at hand, taking also into account the kind of images/videos of the target application scenario. For instance, the presence of eyeglasses is not likely to be detected by descriptors that do not consider the head as a distinct body part, or if the image resolution is too low. Note finally that, if a supervised procedure is used for implementing detectors (see below), a set of images of individuals exhibiting each attribute of interest must be collected.

Implementation of detectors

In principle, the implementation of the detector for a given attribute depends also on the kind of the input descriptor. In fact, ad-hoc detectors were considered in [96, 94]. The MCD

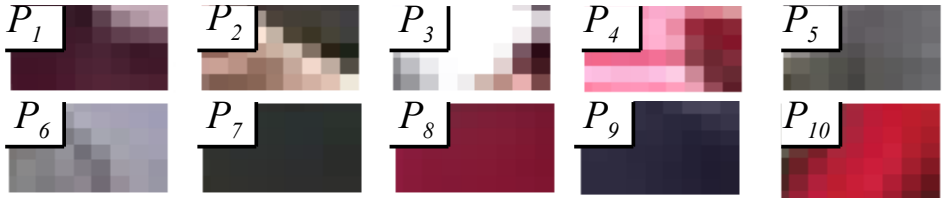

Figure 4.1: Example of the image patches (components of a MCD descriptor) corresponding to different prototypes, obtained from the upper body parts of images of individuals taken from the VIPER data set (see Sect. 4.2).

framework suggests an approach independent on the specific descriptor, instead. Our intuition is that the clothing characteristics that can be detected by using a given appearance descriptor, according to its low-level features and subdivision into parts, can be encoded by one or more visual prototypes. For instance, Fig. 4.1 shows image patches extracted from the upper body parts of individuals taken from the data set of Sect. 4.2, using the MCD implementation of [82]. Each patch corresponds to a single component of a different prototype, the one closest to the cluster centroid. If components are described using colour features, then the descriptors of individuals wearing a red shirt can be expected to exhibit a high similarity to prototypes P_8, P_{10}, and perhaps P_4, and a lower similarity to the other ones. Similarly, the descriptor of an individual wearing a white shirt should exhibit a high similarity to P_3, and possibly P_6 and P_5, if some parts of the shirt are in shade. This suggests that MCD descriptors can be conveniently exploited as input features of attribute detectors, independently on the underlying appearance descriptor.

Since responding to each textual query is a retrieval/ranking problem, as well as re-identification with image queries, we want the result of a query to be an ordered sequence of images, ranked with respect to their relevance, i.e., to the likelihood that the corresponding characteristic is present. Accordingly, detectors should output a real-valued score rather than a crisp decision. If a crisp decision is needed, a threshold can be set according to a suitable, application-dependent criterion (e.g., one can threshold the detector scores, or choose to retrieve only the top-N images according to their scores, for a given N).

Since the MCD descriptor is a fixed-size vector of scalars, the problem of implementing a detector for any given attribute can be seen as a supervised, binary classification problem in a feature space made up of dissimilarity values between an image descriptor and the prototypes: it consists of recognising whether the corresponding attribute is present or not in an input image, as in [53]. The training set can be obtained from a gallery of images of individuals, labelled according to the presence or absence of the considered attribute. Any classification algorithm that outputs real-values scores can thus be used, like support vector machines and neural networks.

Note finally that, since each basic query can be related to a subset of body parts, only the corresponding components of the MCD dissimilarity vector can be used as the input of the corresponding detector. For instance, in the case of a body subdivision into upper (torso and arms) and lower (legs) parts, the descriptor of the upper part does not carry any useful information about an attribute like “red trousers/skirt”. This easily allows detectors to be implemented using the relevant feature subsets only.

Processing basic and complex queries

The answer to each basic query is obtained by running the detector of the corresponding attribute through all the images at hand, and returning a list of images ranked according to the detectors' scores. Denoting the set of basic queries as $\{q_1, q_2, \ldots\}$, complex queries can be formulated by combining any subset of basic ones with Boolean operators. For instance, if q_1, q_2 and q_3 denote respectively the attributes "red shirt", "blue trousers/skirt" and "black trousers/skirt", then the query "individual wearing blue or black trousers, and a red shirt", is encoded by

$$Q = q_1 \wedge (q_2 \vee q_3)\,. \tag{4.1}$$

Processing a complex query is not straightforward, instead. At least two approaches can be followed. One approach consists of running first each component basic query, independently on each other, and then converting their answers into crisp logic values (True or False). In this way, the semantic of a complex query Q becomes the same of classical logic, and its answer can be simply obtained by combining the sets of images retrieved by each basic query, through the set operators corresponding to the logic operators in Q. For instance, in the above example the answer to the complex query (4.1) would be obtained by the intersection between the sets of images retrieved by q_1, and the union of the sets of images retrieved by q_2 and q_3. To convert the answers of a basic query into a crisp logic value, one could set a threshold on the output of the corresponding detector, so that the associated attribute is deemed present if the score is above the threshold, and absent otherwise. For instance, the threshold can be set according to a desired precision-recall trade-off. Alternatively, one can choose to retrieve only the top-N ranked images according to their scores, for a given N. The most suitable criterion is obviously application-dependent.

Another possibility is to use a fuzzy logic approach. Under this viewpoint, the detector score produced for each basic query can be considered as a fuzzy truth value (provided it is properly rescaled into $[0,1]$), corresponding to the statement that the individual in the input image exhibits the associated attribute. In other words, the detector can be seen as implementing the fuzzy membership function for such a statement. Accordingly, the answer to a complex query is obtained by combining the score values of basic queries with fuzzy logic operators, and by returning the list of input images ranked according to the resulting combined score. For instance, implementing the fuzzy-AND and fuzzy-OR respectively with the minimum and maximum operators, the score of an input image $\mathbf{I}$ for the complex query (4.1) is obtained as

$$\min\{s_1(\mathbf{I}), \max\{s_2(\mathbf{I}), s_3(\mathbf{I})\}\}\,, \tag{4.2}$$

where $s_i(\cdot)$ denotes the score of the detector associated to the basic query q_i.

4.2 Experimental evaluation

In this section we describe a possible implementation of our semantic retrieval of pedestrians approach, and its empirical evaluation on a hand labellSed benchmark data set for person re-identification.

4.2.1 Implementation

Appearance descriptors

We considered two different descriptors previously proposed for person re-identification tasks. The first one is the SDALF descriptor [32]. It subdivides body into torso and legs, and represents each part with three local features: an HSV colour histogram, the "Maximally Stable Colour Regions", and the "Recurrent Highly Structured Patches". We used the first two features, that are related to colour information. The second one is the descriptor we proposed in [85]. It subdivides the body as in SDALF, and represents each part with the HSV colour histograms of a bag of randomly extracted 80 image patches. We also used a variant in which the body is subdivided using a pictorial structure [2] into nine parts: the torso and the upper and lower part of every limb. We then obtained a MCD descriptor from each of the above multiple part-multiple component descriptors, as explained in Sect. 2.2. We will denote them respectively by MCD_2, MCD_1 and MCD_3.

To construct a MCD descriptor, prototypes have to be extracted first, from a given image gallery of individuals. This can be made in two different ways, depending on the application scenario. In a scenario like the off-line, forensic analysis of a fixed data set of images (or videos), the images one wants to search can be entirely available beforehand. In this case, prototypes can be conveniently extracted from all such images (we remind the reader that this step is totally unsupervised, and thus does not require any manual labelling of images). In other scenarios, one should instead extract prototypes off-line from a design gallery, and then use them to compute the dissimilarity representations of different pedestrian images at operation phase. In this case, the retrieval performance of the proposed method may be affected by the representativeness of the design gallery with respect to images processed at operation phase. We chose to carry out our experiments under the latter scenario, which can be the most challenging one. Nevertheless, the experimental evidences reported in [82] suggest that, if a design gallery containing a wide range of different clothing characteristics is used, the prototypes are likely to be representative even of a different image gallery. Since we subdivided the considered data set into a training set, to construct attribute detectors, and a testing set, to evaluate the performance of detectors (see below), prototypes were extracted from training images. We used to this aim the two stage clustering procedure of [82]. Note that in MCD_2, two different sets of prototypes were created, one for each kind of local features. We used the K-th Hausdorff for computing dissimilarities, with $K = 10$.

Data set

We used the VIPER data set [43], which is one of the benchmarks for re-identification tasks. It is made up of 1264 images of 632 pedestrians, exhibiting different lighting conditions and pose variations (see the examples in the figures below). For each pedestrian, two images taken from two different cameras with non-overlapping views are present. The image size if 48×128 pixels.

Attributes choice

All the above descriptors enable queries related to clothing colour. In particular, MCD_1 and MCD_2 enable queries related to upper or lower body, like "individual wearing a white shirt".

MCD_3 should also enable more specific queries, like "short sleeves", since the corresponding attribute could be detected by the presence of skin-like colour in lower arms. Accordingly, we identified fifteen attributes corresponding to clothing characteristics that can be detected by the above descriptors, and are also present in at least 5% images of the whole VIPER data set, so that it was possible to build a training set for implementing the corresponding detector. The chosen attributes are related to the colours of the upper and lower body parts, and to the presence of short sleeves/trousers/skirts. They are reported in Table 4.1, where the corresponding number of positive samples is shown between brackets.

We then manually tagged all VIPER images separately for each attribute, using the following criteria. We labelled an image as positive for attributes related to a colour of upper and lower clothing, it that colour appeared approximately in at least one-third of the considered body part. In the case of short sleeves/trousers/skirt, a positive label was given if at least half of the limb was visible. Note that for some images the colour of the upper or lower body garments, or the presence of short sleeves or short trousers/skirts, was not clear, due for instance to low image quality, occlusions, or shadows. We discarded these images both from training and testing samples. This is the reason why the number of labelled samples reported in Table 4.1 for each attribute is lower than the size of the VIPER data set. We point out that this is a correct procedure for training samples, while in a real application such kind of images may appear among testing ones. In Fig. 4.2 we show one positive sample for each attribute, and some of the discarded images. All the tagged VIPER images are available at http://tinyurl.com/peoplesearch-pralab.

Experimental setup

For each attribute, we randomly subdivided the images of the VIPER data set into a training set and a testing set of identical size. We used stratified sampling, so that the same number of positive samples were present in the training and testing set. We then extracted the prototypes from training images. Different numbers of prototypes were considered, ranging from 5 to 300. The results reported in the following refer to 200 prototypes for MCD_1, and 100 prototypes for MCD_2 and MCD_3. We will then discuss how the number of prototypes affect the performance. The detector of each attribute was implemented using a two-class support vector machine (SVM) classifier with radial basis function (RBF) kernel. We used the LIBSVM software [20] for training SVMs. We set the C parameter of the learning algorithm to the LIBSVM default values. Since for most attributes positive examples were much less than negative ones, we set the misclassification cost of the former ten times higher than for the latter. The γ parameter of the RBF kernel was set to 100. We then evaluated the performance of each detector on testing images. We repeated this procedure for ten times, and report in the following the average results over the ten runs. Since we are dealing with a retrieval task, we used the precision-recall (P-R) curve to evaluate the performance of each detector (i.e., the retrieval performance for each basic query). It was computed by thresholding the detectors' (SVM) outputs. We also considered some complex queries, and processed them using the fuzzy logic approach described in Sect. 4.1.

Table 4.1: The fifteen attributes considered in our experiments, related to upper body (top nine rows), and lower body (last six rows). #positive denotes the number of images labelled as exhibiting the corresponding attribute; #negative denotes the overall number of images labelled either as positive or negative (ambiguous images were discarded). The three rightmost columns report the break-even point (BEP) of the precision-recall curves attained on testing images by the considered descriptors, averaged over the ten runs of the experiments. For each attribute, the highest BEP over the three descriptors is shown in bold.

Attributes (#positive/#labelled)	MCD_1	MCD_2	MCD_3
red shirt (73/904)	**0.69**	0.66	0.62
blue shirt (84/904)	**0.48**	0.42	**0.48**
pink shirt (42/904)	0.50	**0.51**	0.44
white shirt (277/904)	0.71	0.73	**0.77**
black shirt (298/904)	0.73	0.68	**0.74**
green shirt (72/904)	0.57	0.49	**0.59**
grey shirt (70/904)	**0.38**	0.30	0.28
brown shirt (71/904)	**0.46**	0.37	0.38
short sleeves (382/1190)	0.54	0.51	**0.60**
blue trousers/skirt (568/978)	0.87	0.85	**0.90**
white trousers/skirt (112/978)	**0.68**	0.58	0.63
black trousers/skirt (178/978)	0.68	0.64	**0.75**
grey trousers/skirt (52/978)	**0.30**	0.18	0.29
brown trousers/skirt (40/978)	**0.62**	0.33	0.49
short trousers/skirt (129/1197)	0.33	0.41	**0.58**

4.2.2 Experimental results

For each of the considered attributes, in Fig. 4.3 we show one example of a clearly related prototype (obtained with MCD_1). This example supports our intuition that prototypes extracted by MCD can also encode high-level visual characteristics of clothing appearance, even though their extraction is totally unsupervised.

The P-R curves of each basic query, for the three considered descriptors, are reported in Figs. 4.4 and 4.5. Table 4.1 also shows the average break-even point (BEP) of each curve, which is the point where precision equals recall. An example of the ten top-ranked images for five out the fifteen basic queries is shown in Fig. 4.6. The retrieval performance depends on the attribute, and on the underlying appearance descriptor that was used to build the MCD descriptor. In general, better performances were attained for attributes with a larger number of positive examples (which is reported in Table 4.1). For instance, a BEP of about 0.5 was attained for the "blue shirt" attribute, which has 84 positive samples, while a BEP between 0.85 and 0.90 (depending on the descriptor) was attained for "blue trousers/skirt", which has 568 positive samples. The retrieval performance for "red shirt" was very good instead, even though the positive samples were only 79. The reason is that the red colour is well separated from the other ones in the HSV space, which is used by all the considered descriptors (note that we did not consider the "red trousers/skirt", since only a few positive examples were available in the VIPER data set). The performance was rather low for attributes related to grey, brown and pink colours (except for brown trousers/skirt and MCD_2), not only be-

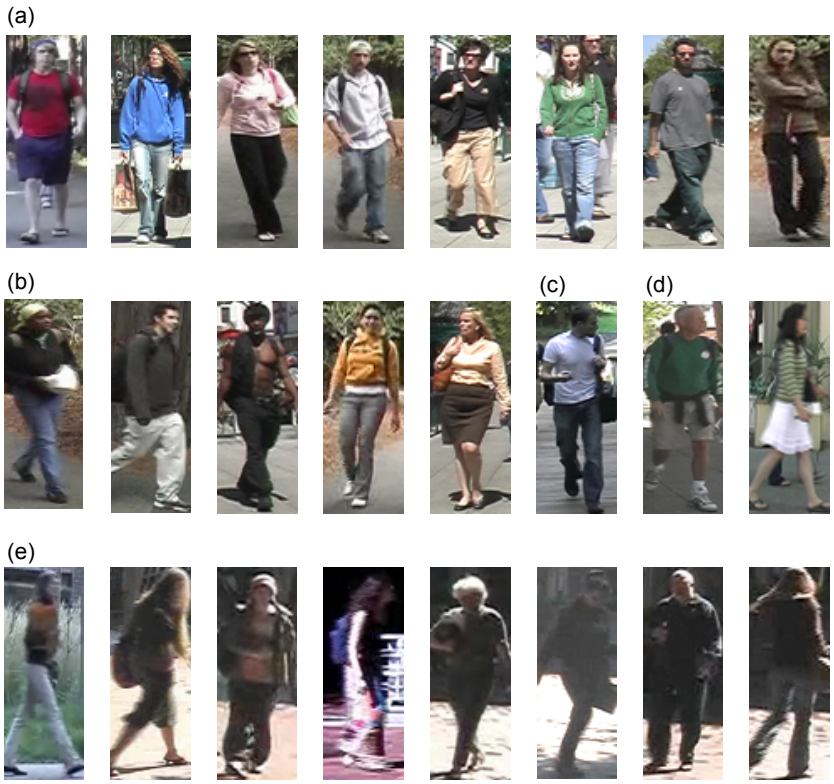

Figure 4.2: Examples of images taken from the VIPER data set. (a)–(d): positive examples for attributes related to (a) upper body clothing colours (from left to right: red, blue, pink, white, black, green, grey and brown shirt); (b) lower body clothing colours (blue, white, black, grey, brown trousers/skirt); (c) short sleeves; (d) short trousers/skirt. (e) Examples of ambiguous images discarded from the data set because of occlusions, shadows or low quality.

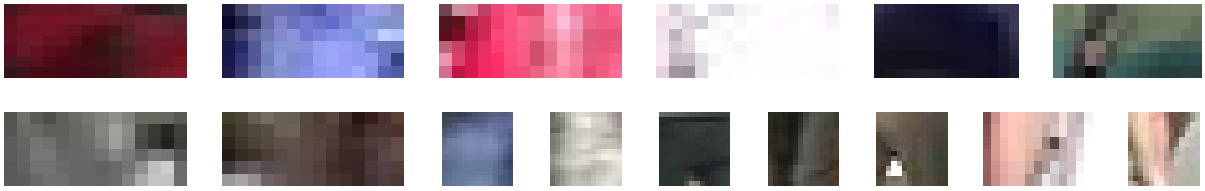

Figure 4.3: Examples of prototypes obtained using the MCD_1 descriptor. Each one can be related to one of the considered attributes. From top to bottom, and from left to right: red, blue, pink, white, black, green, grey and brown shirt (prototypes obtained from the upper body); blue, white, black, grey, brown trousers, skirts (prototypes obtained from the lower body); short sleeves (upper body), short trousers/skirt (lower body).

cause of the small number of positive samples, but also because these colours are not well discriminated by the HSV space.

Consider now the attributes “short sleeves” and “short trousers/skirt”, whose detection relies on skin colour in arms or legs. As pointed out in Sect. 4.1, MCD_3 was likely to attain the best performance on such attributes, due to its more refined body subdivision. This is confirmed by the corresponding P-R curves (see the last two plots of Fig 4.5). Moreover, the part detection technique used by both MCD_1 and MCD_2 (see [32]) turns out to produce a

better quality mask for the upper body than for the lower body, and this is the reason why a higher performance is attained by MCD_1 and MCD_2 for "short sleeves" than for "short trousers/skirt".

We also considered some examples of complex queries. Here we report some results for "white shirt and blue trousers", and "white shirt and short sleeves". We processed them using the fuzzy logic approach, and thus computed for both queries the minimum of the score provided by the detectors of the two embedded basic queries. The ten images exhibiting the highest combined score are shown in Fig. 4.7. Although these results are very limited, they provide some evidence that a fuzzy logic approach can be a convenient one to process complex queries.

We finally evaluated how the retrieval performance is affected by the number of prototypes. We observed that the performance initially grows as the number of prototypes increases, then reaches a nearly stable value. Such a value was around 100 for MCD_1 and MCD_3, and 200 for MCD_1, with small variations depending on the basic query. This behaviour can be easily explained: once the number of prototypes large enough that most of the distinctive visual characteristics are captured by different clusters, increasing the number of prototypes has only the effect of splitting some of the previous clusters into two or more similar ones. Consequently, no additional information is encoded by new prototypes. On the contrary, a too high number of prototypes may increase the risk of over-fitting, when dissimilarity vectors are used as features of classification algorithms.

4.3 Conclusions

We proposed a general method for retrieving images of pedestrians with queries related to clothing appearance attributes, using any multiple part-multiple component descriptor developed for person re-identification tasks. This allows us to extend re-identification systems, by including a search functionality based on *textual* queries.

An interesting direction for further research is to extend our approach to 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 as well. In this case, a bag of dissimilarity vectors coming from different frames is available for each tracked individual, instead of a single one. Accordingly, a Multiple Instance Learning approach [27] could be used to build the detectors.

Another interesting issue is using Natural Language Processing techniques [49] to automatically encode an original textual description of an individual of interest into a Boolean combination of the available basic queries.

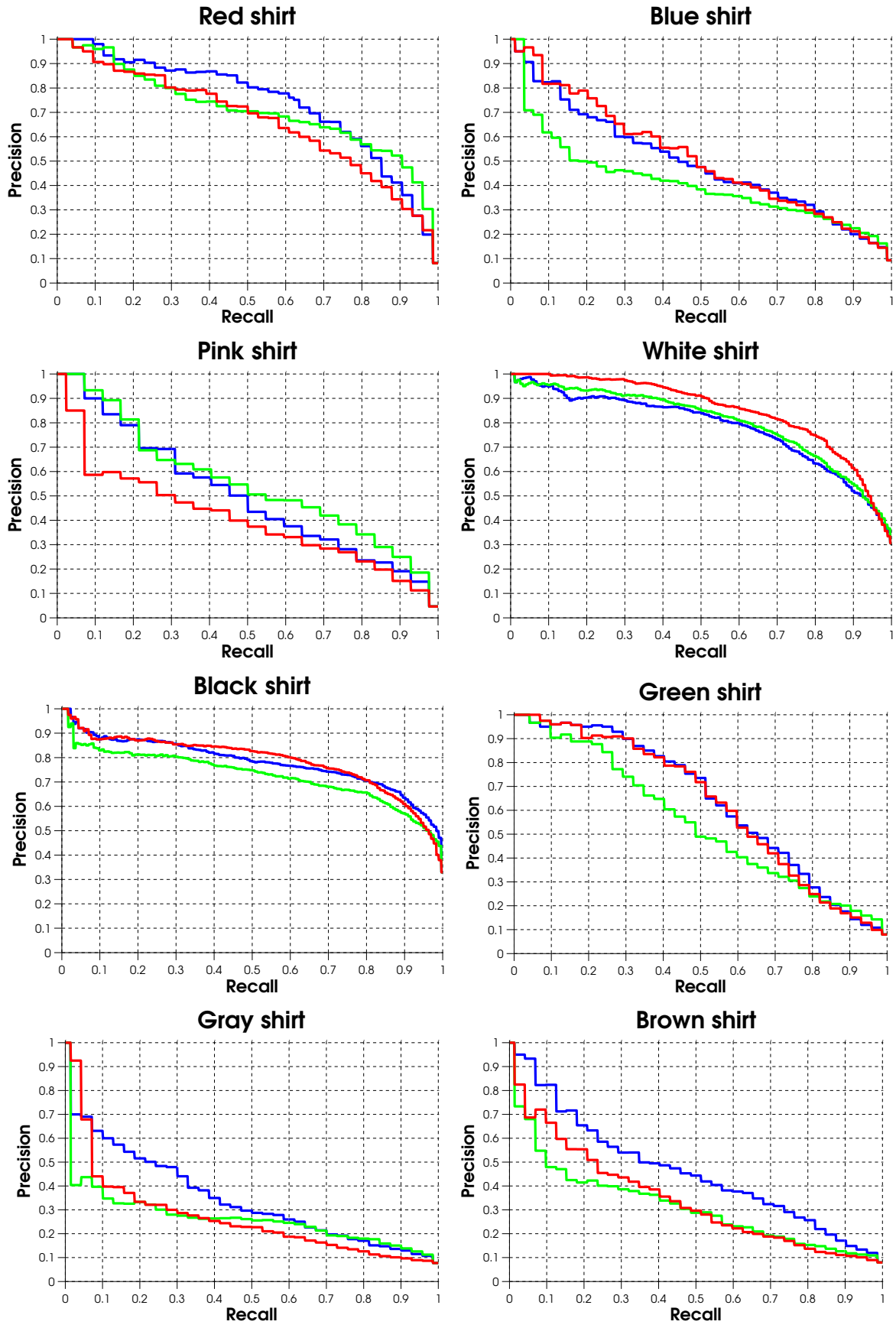

Figure 4.4: Average P-R curves for the eight basic queries related to the clothing colours of the upper body. Blue: MCD_1; green: MCD_2; red: MCD_3.

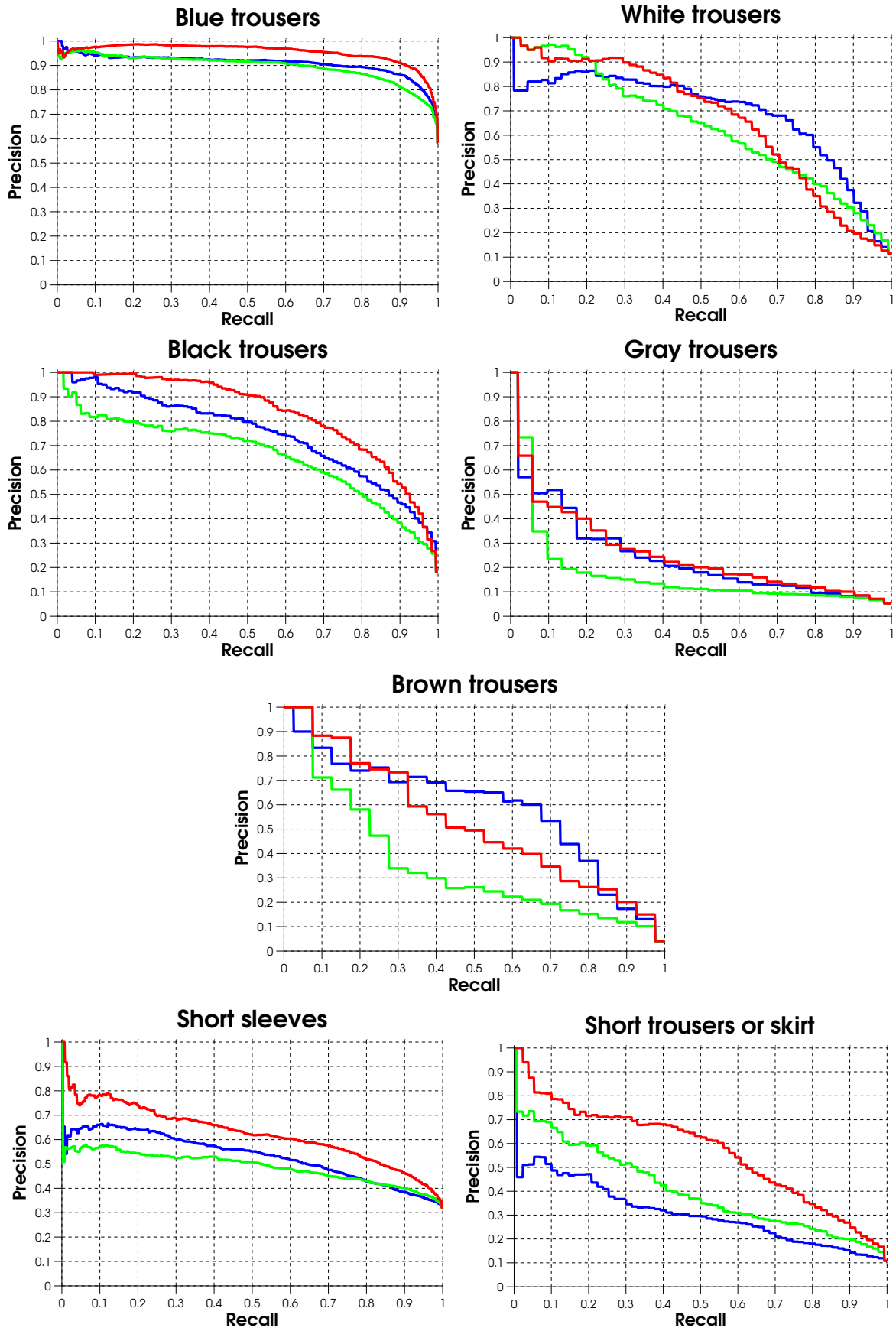

Figure 4.5: Average P-R curves for the five basic queries related to the clothing colours of the lower body (top five plots), and to short sleeves and short trousers/skirts. Blue: MCD_1; green: MCD_2; red: MCD_3.

Figure 4.6: The top ten images retrieved by MCD_2, for the queries "red shirt" (a), "pink shirt" (b), "white shirt" (c), "green shirt" (d) and "white trousers" (e), sorted from left to right for decreasing values of the score provided by the corresponding detectors. Non-relevant images are highlighted in red.

Figure 4.7: The top ten images retrieved by MCD_3, for the queries "white shirt and blue trousers" (a) and "white shirt and short sleeves" (b). The images are sorted from left to right, for decreasing values of the score computed using a fuzzy logic approach, as the minimum of the scores produced by the detectors of the embedded basic queries. Non-relevant images are highlighted in red.

Chapter 5

Semantic retrieval of pedestrians via deep representations

Person re-identification consists of searching for an individual of interest in video sequences acquired by a camera network, using an image of that individual as a query. Here we consider a related task, which consists of searching images of subjects that match a semantic description of their appearance, given by a combination of predefined attributes. This allows for automatic image annotation of pedestrians, useful to organize and locate images of interest from the video recordings acquired by a surveillance camera network.

Semantic image retrieval can be useful also in applications such as forensic video analysis, where the query can be obtained from a eyewitness report. In wider terms, learning good representations of the data in this specific domain knowledge, can be useful to help designing more general representations of pedestrians, that can be used for other tasks such as person re-identification.

In this chapter we propose a general method for implementing semantic image retrieval using deep architectures, that are formed by the composition of multiple non-linear transformations, taking advantage of the multiple part appearance representation of the pedestrian image. For each part, formerly head, torso, legs and shoes, we train a convolutional neural network aimed to detect the semantic attributes relative to that particular part. For example, the network trained with head images is designed to detect attributes such as the color of the hair, while the torso images are used as examples for detecting attributes such as clothing type, color and the presence of accessories.

Whereas the decision of the network architecture relies on trial and error experiments to choose the proper parameters, the loss function plays a key role in modelling what is the purpose of the training procedure. Therefore, we investigate different loss functions to find what suites the best for the task at hand, which is a multi-label annotation problem. We selected different approaches from the state of the art, and also did some modification to take into advantage of the peculiarity of the domain we are considering. In contrast with the state of the art, we exploit the information about the contemporary presence of attributes, to take advantage of the correlation of the data, peculiar of this specific domain. For instance, characteristics such as "‘white hair” are mutually exclusive with "‘black hair”, while concepts such as "‘accessory hair band” are related to "‘long hair”.

Finally, we tested the proposed framework on a hand-labelled ensemble of 19000 images

Datasets	#Images	Camera angle	View point	Illumination	Resolution	Scene
3DPeS	1012	high	varying	varying	from 31x100 to 236 x 178	outdoor
CAVIAR4REID	1220	ground	varying	low	from 17x39 to 72x141	outdoor
CUHK	4563	high	varying	varying	80x160	indoor
GRID	1275	varying	frontal & back	low	from 29x67 to 169x365	indoor
i-LIDS	477	medium	back	high	from 32x76 to 115x294	outdoor
MIT	888	ground	back	high	64x128	outdoor
PRID	1134	high	profile	low	64x128	outdoor
SARC3D	200	medium	varying	varying	from 54x187 to 150x307	outdoor
TownCentre	6967	medium	varying	medium	from 44x109 to 148x332	outdoor
VIPeR	1264	ground	varying	varying	48x128	outdoor
Total = PETA	**19000**	**varting**	**varying**	**varying**	**varying**	**varying**

Table 5.1: Characteristics of the PETA dataset [26]

of pedestrians acquired from video surveillance recordings, taken from several benchmark datasets for the person re-identification task. We show the results of extensive experiments in terms of precision-recall curves, able to well characterize both the accuracy of the retrieved results, and the error in terms of subjects left out from the query response.

In section 5.1 we explain in detail our approach, in 5.2 the results of experimental evaluation and 6 concludes the chapter with some insights on future work.

5.1 Our approach

In this section we explain our newest approach for semantic retrieval of pedestrian images. We start presenting the evaluation benchmark, explaining its peculiarities in section 5.1.1. In section 5.1.2 and 5.1.3 we progress explaining what pre-processing and data augmentation is needed in order to take into advantage of the multiple-part composition of the human body and to avoid over-fitting of the networks presented in section 5.1.4. Finally, in section 5.1.5 are explained the loss functions we exploited for the multi-label annotation problem. In respect to the state of the art, the contribution of this work can be summarized in three points: (i) instead of using the whole image to train the attribute detectors, we employ a deep neural network for each body part (ii) we propose a new loss function aimed at exploiting the correlation among the attributes given as ground truth (iii) we perform extensive experiments to asset both the detection performance with images of pedestrians acquired from the same cameras/locations used in the training phase, and testing the generalization properties using images taken from a totally different dataset.

5.1.1 The dataset

Composition of PEdesTrian Attribute (PETA) [26] is a dataset of 19000 labeled images of 8705 pedestrians, with resolution ranging from 17-by-39 to 169-by-365 pixels, annotated with 61 binary and 4 multi-class attributes. The images are taken from several benchmark datasets for person re-identification: VIPeR [42], 3DPes [5], CAVIAR4REID [17], CUHK [58], GRID [60], i-LIDS [48], MIT [76], PRID [47], SARC3D [6] and TownCentre [12]. In figure 5.1 is depicted the composition of the dataset and some sample image. In table 5.1 are indicated some characteristics such as the number of images for each subset, the camera angle, viewpoint, illumination, resolution and scene (indoor/outdoor).

Figure 5.1: Composition of the PETA dataset [26] and some sample images

We point out that the differences in camera angle, view point, illumination, resolution and scene, allow for a more robust detection of attributes, since the avoidance of these nuances will be learned directly from the data. This is important especially in terms of representation learning of the pedestrian image, since this allows for entangling and hiding more or less the different explanatory factors of variation behind the data [13].

5.1.2 Pre-processing

Pre-processing is an essential step for detecting attributes since it allows, from the video recordings, to access the pedestrian image and its segmentation in body parts. The first process consists of extracting the bounding boxes of pedestrians from the video surveillance recordings. Since they are already provided by the authors of the datasets composing PETA, the reader can refer on their work for the person detection task. In this section we focus on background subtraction, which consists on considering just the pixels of the image appertaining to the silhouette of the individual, and body part segmentation, which consists of establishing which pixels appertain to the head, torso and legs.

We employed the method of [62], which proposed a Deep Decompositional Network for parsing pedestrian images into semantic regions, where the subjects can be heavily occluded. We chose this method because targeted to pedestrian images acquired at far distance, and because of the low computational complexity. In figure 5.2 are some results of the segmentation on the Penn-Fudan data set [98]. We point out that these images are not overlapping with the PETA dataset used for attribute detection. To feed the pedestrians images on the trained deep decompositional network we resized all the images to 160-by-60 pixels. Once the segmentation is obtained, we resize each body part to a standard size in order to fit into a dedicate convolutional neural network. For the head part we resized the images to 32-by-32 pixels, the torso to 64-by-64 pixels and the legs to 128-by-64 pixels. This choice has been done heuristically in order to keep the aspect ratio and a number of pixels equal to some power of two.

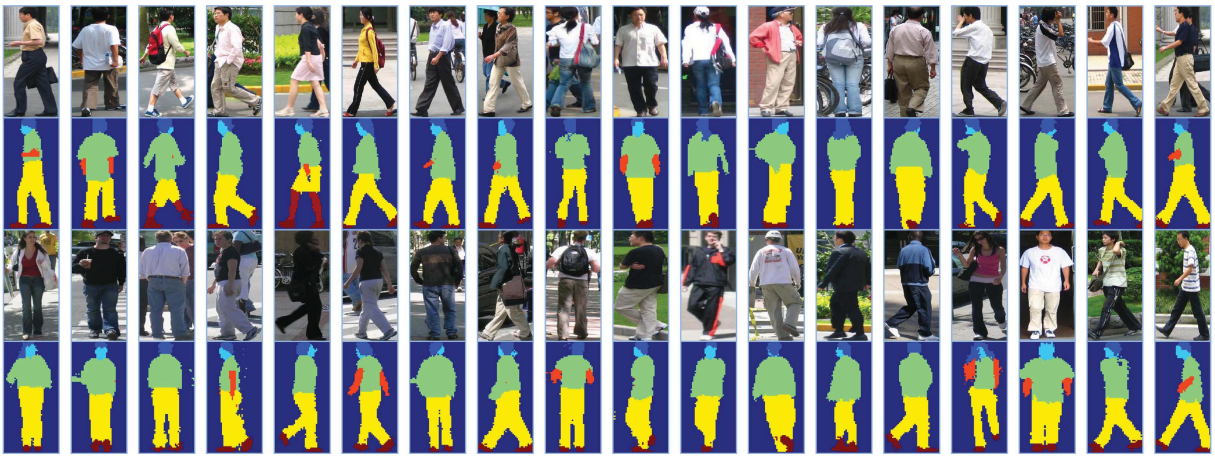

Figure 5.2: Some results of image segmentation with the method proposed in [62]

5.1.3 Data augmentation

Data augmentation is a technique widely used in convolutional neural networks, being able to avoid over-fitting and improve robustness. In particular, label-preserving transformations have shown to be very effective [52] in regularizing the deep architectures and boosting performances. In this work we employed horizontal flipping of the pedestrian image, and a technique that consists on altering the intensity of color pixels, usually known as fancy pca, first employed in [52].

The intensities of the RGB channels of the training pedestrian images are summed to a quantity multiple of the principal component. Denoting with $I_{x,y} = \{I^R_{x,y}, I^G_{x,y}, I^B_{x,y}\}$ each image of the training set, we add a quantity proportional to the eigenvalues λ_i of the 3×3 covariance matrix of the RGB pixel values:

$$[\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T \tag{5.1}$$

where $\mathbf{p_i}$ are the eigenvectors and α_i are random variables drawn from a Gaussian distribution with mean zero and standard deviation 0.1. As claimed by [52], this technique is able to take into advantage of an important property of natural images, namely, that object identity is invariant to changes in the intensity and color of the illumination. Differently from [52], we avoided making random crops of the images to augment the dataset. Since the training images are the result of person detection algorithms, a crop would probably remove an essential part of the image useful to recognize the considered attributes.

In the experiments we increased the dataset size by a factor of 10, generating four color alterations of the original image and the horizontal flipped one. We did this just for the training images, while for validation and test we considered the original set.

5.1.4 The architecture

In this section is described the chosennetwork architecture for the multi-label annotation problem. We employed a different network for each body part, therefore there are three networks, taking as input the resized cropped image as explained in 5.1.2. In figure 5.3 an illustration of the network corresponding to the head part.

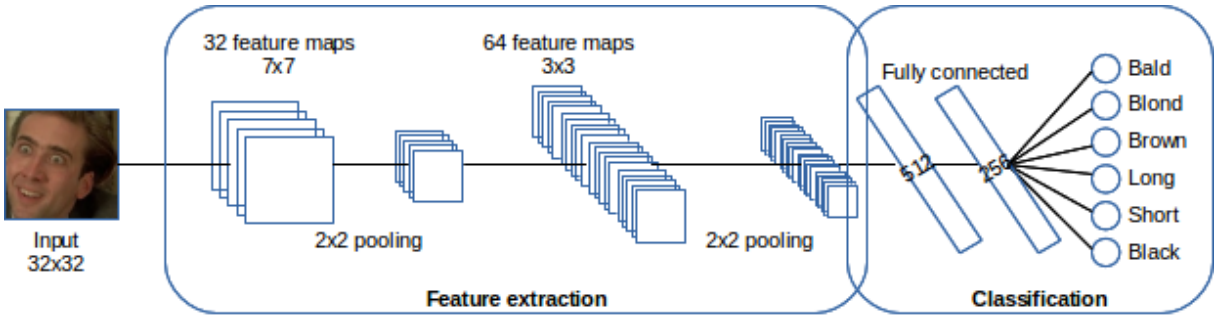

Figure 5.3: Network architecture

Each network consists of two convolutional and two fully-connected layers. The first convolutional layer consists of 32 filters of size 13×13, a rectifying linear unit and a 2×2 max-pooling. For the head part we decreased the size of filters to 7×7 because smaller than the torso and legs parts. The second convolutional layer has 64 filters of size 3×3, and is followed by another rectifying linear unit and a 2×2 max-pooling. The final two fully connected layers come with dropout [91] regularization with probability 0.6. The output layer has a number of neurons equal to the number of attributes for that particular part, and the non-linearities are sigmoids since the labels are $[0,1]$ values indicating the absence or presence of each attribute in the input image. We trained the networks using stochastic gradient descent with a batch size of 200 examples, momentum of 0.9, learning rate of 0.0005 and L2 normalization.

5.1.5 Multi-label losses

The main focus of this work consists on investigating the different objective functions that can model the multi-label annotation problem at hand. The first loss we considered is the classical binary cross-entropy, explained in section 5.1.5. In section 5.1.5 is presented the pairwise ranking loss of [50], that directly models the annotation problem. We implemented a modification in order to exploit the co-occurrence of the attributes, factorizing the joint probabilities into the model, in accordance with the inference algorithm presented by [36].

Formally, we start considering a representation of images $x \in \Re^D$ and a representation of annotations $i \in \mathrm{Y} = \{1, \ldots, \mathscr{Y}\}$, taken from the tags dictionary. We assume that each image has multiple labels, and that we can form a label vector $\mathbf{y} \in \Re^{1 \times c}$ where $y_j = 1$ means the presence of a label and $y_j = 0$ means absence of a label for an image x. The task consists on ranking labels $i \in \mathscr{Y}$ given an example x. The n labelled pairs $(x, \mathbf{y})$ provided in the training phase, have a set of c_+ positive annotations and a set of c_- negative annotations so that $c = c_+ + c_-$, where c is the number of tags. We assume that we have a set of images $\mathbf{x}$ and denote the output of the convolutional network as $f(\cdot)$.

Binary cross-entropy

The binary cross-entropy [24] is widely used to define the loss function in many machine learning and optimization tasks. Given a single label y_i, and the output of the network $f(x)$, we can use the cross entropy to get a measure of similarity between the predicted probability and the ground truth as:

$$err\left(f(x), y\right) = -y \log\left(f(x)\right) - (1-y) \log\left(1 - f(x)\right) \tag{5.2}$$

In the multi-label scenario we have c different labels for each training image x, therefore we can consider as cost function the average of the cross-entropies relative to each tag:

$$\overline{err}\left(f(x),y\right) = -\frac{1}{n\cdot c}\sum_{i=1}^{n}\sum_{j=1}^{c^+} y_{ij}\log\left(f_j(x_i)\right) - \frac{1}{n\cdot c}\sum_{i=1}^{n}\sum_{k=1}^{c^-}(1-y_{ik})\log\left(1-f_k(x_i)\right) \tag{5.3}$$

In this work we got better results minimizing the entropy between the predictions and the ground-truth probabilities. The ground truth probabilities $\overline{p_{ij}}$ can be evaluated by counting the number of occurrences of the correspondent attribute in the training set, and dividing by the number of examples n:

$$\overline{err}\left(f(x),y\right) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c^+}\overline{p_{ij}}\log\left(f_j(x_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c^-}(1-\overline{p_{ik}})\log\left(1-f_k(x_i)\right) \tag{5.4}$$

Pairwise Ranking

The second loss we considered is the pairwise ranking loss, first introduced by [50] and modified by [41] to take into account of multiple labels. The function has the purpose of ranking the positive labels to have higher scores than negative ones:

$$\overline{err}\left(f(x),y\right) = \sum_{i=1}^{n}\sum_{j=1}^{c^+}\sum_{k=1}^{c^-}\max\left(0,1-f_j(x_i)+f_k(x_i)\right) \tag{5.5}$$

The advantage of this loss is that it is aimed to optimize the area under the ROC curve (AUC) for each attribute in the set. In this work, we take also into advantage of the correlation between output labels, that can be used to factor the joint probabilities in order to improve predictions of the attributes. In addition to the individual probabilities, we can model the probability of labels occurring together using the following inference algorithm [36]:

$$P\left(y|x\right) = \prod_{j} P\left(y_j|x\right)\prod_{k,l} P\left(y_l|y_h\right)^{\alpha} \tag{5.6}$$

where $l,h\in\{i|y_i=1\}$ and $0\leq\alpha\leq 1$. $P\left(y_j^{(m)}|x^{(m)}\right)$ is the probability of the independent binary model for each label i of the set of attributes, given the image x. This can be considered as the output of the convolutional neural network $f(x)$. The co-occurrence is represented as $P\left(y_l^{(m)}|y_h^{(m)}\right)$ and can be extracted from the covariance matrix of the ground-truth labels in the training set. This quantity does not depend on the input images, but it is computed once as a prior for the entire dataset. These probabilities are scaled by a factor α since co-occurrence of labels would also require higher order moments that we discarded because of computational overhead [36].

It is easy to see that for a logarithmic loss function, this inference algorithm gives no contribution since the gradient in respect to the inputs is zero and it would not back-propagate:

$$\log P\left(y|x\right) = \sum_{j}\log P\left(y_j|x\right) + \alpha\sum_{l,h}\log P\left(y_l|y_h\right) \tag{5.7}$$

being the second addend constant in respect to x, it would give no additional contribution to the binary cross-entropy loss. Therefore we employed this model just for the pairwise loss:

$$\overline{err}\left(f(x),y\right) = \sum_{i=1}^{n}\sum_{j=1}^{c^+}\sum_{k=1}^{c^-}\max\left(0,1-f_j(x_i)\cdot\prod_{l,h}P\left(y_l|y_h\right)^{\alpha}+f_k(x_i)\right) \tag{5.8}$$

For this loss we replaced the network non-linearities to leaky rectifier units [65] and initialized the bias to 0.5 instad of 1 in order to make the training algorithm to work. We point out that using this loss function, it is possible to take advantage only of the co-occurrence of the set of labels specific to the body part in which the convolutional neural network is trained. In figure 5.4 and 5.5 the correlation of the labels considered for each body part.

5.2 Experimental evaluation

5.2.1 Implementation

The approaches and methods explained in the previous section have been implemented using Theano [9, 14], which is a python library that allows to define network models in a symbolic way and to compile them into C and GPU code automatically in an efficient way. We also used a lightweight library to build and train neural networks in Theano, called Lasagne[1]. The advantages of this library consist in a more easy way to define the network architecture and optimization, keeping the great flexibility of Theano. In particular, we took advantage of the possibility of defining a custom cost function with no need to derive gradients because of symbolic differentiation.

Experiments have been carried by a NVIDIA Digits DevBox [2], that comes with Four TITAN X GPUs with 7 TFlops of single precision, 336.5 GB/s of memory bandwidth, and 12 GB of memory per board. Rather than using multiple gpus to train a single network, we preferred doing multiple experiments on different gpus, in order to try different approaches, tricks and parameters.

5.2.2 Choice of the attributes

Since some attributes are more rarely present then others, a selection from the 104 attributes of the PETA dataset [26] is needed. We discarded the attributes that are present less than in the 1% of the dataset. Additionally we discarded the attributes regarding the shoes part, because it is not always visible in the pedestrians images composing the dataset. In the experiments, for the head part we considered the attributes:

- hair style: short, long, bald
- hair color: brown, black, white, yellow, grey
- accessories: sunglasses, hat, hair band

For the torso part:

- upper body color: orange, yellow, pink, green, purple, brown, red, blue, grey, white, black
- textures: thick and thin stripes,
- clothing: sweater, no/short/long sleeves, suit, jacket, t-shirt

[1] https://github.com/Lasagne/Lasagne
[2] https://developer.nvidia.com/devbox

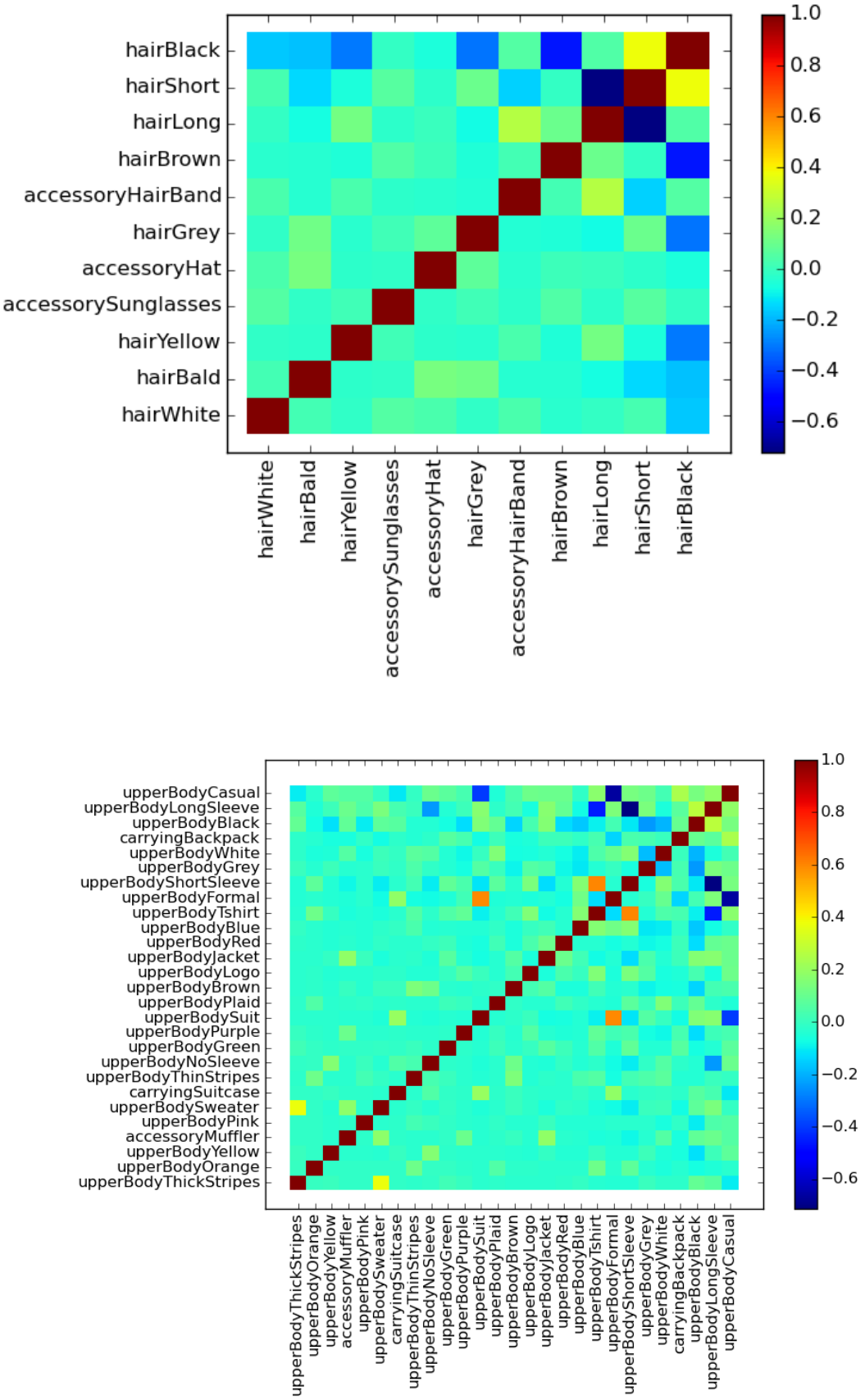

Figure 5.4: Correlation matrices for attributes relative to head and torso body parts

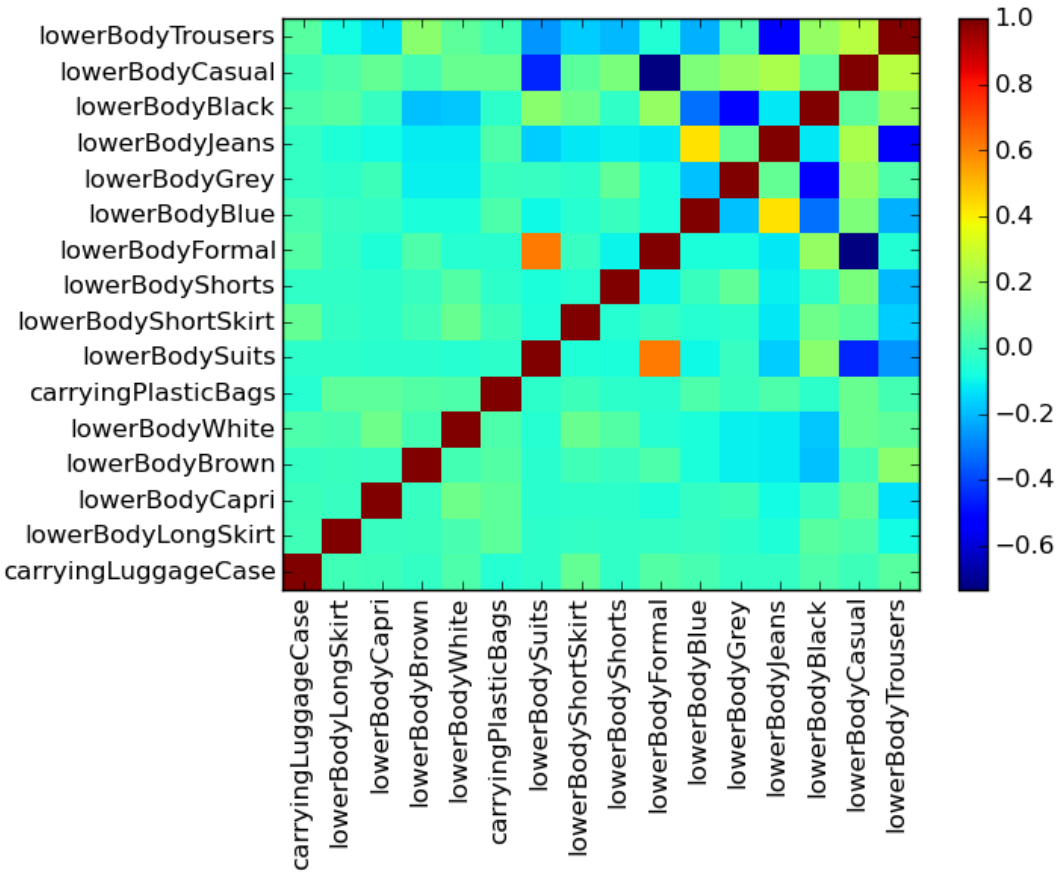

Figure 5.5: Correlation matrices for attributes relative to the legs

- clothing style: casual, formal
- accessories: muffler, plaid, backpack
- logo

For the legs part:

- lower body color: white, brown, blue, grey, black
- clothing: long/short skirt, suits, shorts, capri, jeans, trousers
- clothing style: casual, formal
- accessories: luggage, plastic bags.

5.2.3 Experimental setup

The PETA dataset [26] is composed of different datasets built for person re-identification benchmarking. In this experimental setup we decided to consider all the examples for training, removing the VIPeR [42] dataset in order to use it for test. As a result, we splitted the 8067

remaining individuals in training and validation set: 15902 frames for training and 1834 for validation (depending on the dataset there is a different number of frames for each subject).

Results are expressed in terms of precision-recall curves and their area (AUC). We point out that these performance measures are more suited for such skewed datasets [23], being able to better represent what is happening in terms of precision (how many retrieved subjects present the searched attribute) and recall (how many correct examples have been left out from the query response).

5.2.4 Experimental results

For each of the considered body parts, we show the performance of the proposed approach for both the validation set and VIPeR dataset [42]. Performance on validation set are useful to test the proposed approach in scenarios where the training examples are composed of images coming from the same cameras/location. VIPeR dataset instead, is composed of images coming from cameras/location different from the training set. Therefore these experiments are aimed at testing the ability of the system to generalize to a new domain.

In table 5.2, 5.4 and 5.6 are the performances relative to the validation set and in table 5.3, 5.5 and 5.7 the performance relative to the VIPeR dataset. For each attribute, the first column indicates the performance using the cross-entropy objective function 5.1.5 while the second using the pairwise function 5.1.5.

The retrieval performance depends on the attribute, and on the loss function that was used to train the convolutional neural networks. In general, better performances were obtained for attributes with a larger number of positive examples. The number of positive examples for each attribute is reported in Table 5.8. For instance, an AUC between 0.005 and 0.295 was obtained for the torso thick stripes attribute, which is presented just in 89 of the individuals, while an AUC between 0.587 and 0.745 was obtained for black lower garment, having 4293 positive subjects. The retrieval performance for orange upper garment was very good instead, even though the pedestrians wearing orange shirts were only 106. The performance was rather low for some of the attributes related to the head part not only because of the small number of positive samples, but also because of the low resolution of the part of the image relative to the head. The performance for the torso and legs part instead are generally better, but some detectors have low AUC values because of the ambiguity in the labelling of the relative attributes. For instance, several images are under-saturated and some colors can be confused with gray, black and white. Other sources of error are due to bad segmentation of the pedestrian images, especially when they are carrying objects such as luggages or plastic bags.

Regarding the two loss functions we considered, is evident that the pairwise objective works better for the major part of the attributes. That is, a loss function aimed at capturing the relation between attributes, is able to better capture information such as the mutually exclusivity or high correlation among attributes. For instance the colors for the lower body part are usually mutually exclusive and works way better with the pairwise loss. The same happens for the attribute jeans, which is highly correlated to the color blue that is much more easy to retrieve. We point out that this is a general trending; for some cases it is difficult to explain the performance of some attributes. For instance, usually the weighted cross-entropy seems to work better for the gray color. We think that this may be because being put in relation to black and white labels, this can make easily get in confusion the detectors in respect to a more independent retrieval model. Moreover, since the different pedestrian frames are

Attributes head	Cross-entropy	Pairwise
White	0.062	**0.485**
Bald	0.013	**0.049**
Yellow	0.113	**0.430**
Sunglasses	0.027	**0.237**
Hat	0.011	**0.082**
Grey	**0.324**	0.043
Hair Band	0.088	**0.236**
Brown	0.170	**0.277**
Long	0.365	**0.655**
Short	0.889	**0.935**
Black	**0.963**	0.958

Table 5.2: Performance on validation dataset in terms of the area under curve (AOC) of the precision-recall characteristics for the head part

Attributes head	Cross-entropy	Pairwise
White	0.013	**0.072**
Bald	0.015	**0.403**
Yellow	0.014	**0.054**
Sunglasses	0.221	**0.333**
Hat	0.103	**0.135**
Grey	0.081	**0.133**
Hair Band	0.041	**0.052**
Brown	**0.386**	0.365
Long	0.528	**0.577**
Short	0.648	**0.649**
Black	0.778	**0.779**

Table 5.3: Performance on VIPeR dataset in terms of the area under curve (AOC) of the precision-recall characteristics for the head part

labelled with the frame that better shows the presence of each attribute, in the training set there can be some positive examples where the presence of the attribute is hidden.

In figure 5.6, 5.7 and 5.8 we show a color representation of the learned filters at the first layer of the convolutional networks, trained using the pairwise loss function. Since at the first layer the network is looking directly at the raw pixel data, this weights are more interpretable in respect to the other layers. We can see that the filters relative to the legs part are nice and smooth, while for the head and torso part there is some pixelated filter. This is an indicator that the network hasn't been trained for long enough, or possibly there is overfitting because low regularization. Since these are just preliminary results, in the future we will take this into account in order to possibly get better results in terms of accuracy of the attribute detectors.

Finally, in figure 5.9, 5.10 and 5.11 we show the first ten pedestrian returned by some significant attribute queries.

Attributes torso	**Cross-entropy**	**Pairwise**	**Attributes torso**	**Cross-entropy**	**Pairwise**
Thick Stripes	0.005	**0.295**	Brown	**0.229**	0.186
Orange	0.011	**0.836**	Logo	0.127	**0.220**
Yellow	0.015	**0.216**	Jacket	0.121	**0.127**
Muffler	**0.118**	0.089	Red	0.118	**0.762**
Pink	0.013	**0.393**	Blue	0.822	**0.886**
Sweater	**0.139**	0.099	Tshirt	0.437	**0.608**
Thin Stripes	0.030	**0.066**	Formal	0.152	**0.566**
No Sleeves	0.041	**0.171**	Short Sleeves	0.480	**0.574**
Green	0.015	**0.078**	Grey	**0.582**	0.512
Purple	0.055	**0.080**	White	0.716	**0.743**
Suit	0.091	**0.619**	Backpack	0.322	**0.326**
Plaid	0.092	**0.578**	Black	0.769	**0.802**
Long Sleeves	0.890	**0.938**	Casual	0.909	**0.944**

Table 5.4: Performance on validation set in terms of the area under curve (AOC) of the precision-recall characteristics for the torso part

Attributes torso	**Cross-entropy**	**Pairwise**	**Attributes torso**	**Cross-entropy**	**Pairwise**
Thick Stripes	0.016	**0.112**	Brown	**0.136**	0.128
Orange	0.013	**0.480**	Logo	**0.109**	0.104
Yellow	0.073	**0.284**	Jacket	**0.051**	0.049
Muffler	**0.015**	0.010	Red	0.252	**0.724**
Pink	0.020	**0.152**	Blue	0.453	**0.477**
Sweater	0.020	**0.038**	Tshirt	0.362	**0.453**
Thin Stripes	**0.022**	0.014	Formal	0.028	**0.163**
No Sleeves	0.057	**0.147**	Short Sleeves	0.421	**0.467**
Green	0.091	**0.262**	Grey	**0.337**	0.265
Purple	**0.047**	0.036	White	0.615	**0.618**
Suit	**0.003**	0.002	Backpack	0.400	**0.473**
Plaid	0.036	**0.099**	Black	**0.708**	0.659
Long Sleeves	**0.814**	0.773	Casual	0.974	**0.978**

Table 5.5: Performance on VIPeR dataset in terms of the area under curve (AOC) of the precision-recall characteristics for the torso part

Attributes legs	Cross-entropy	Pairwise	Attributes legs	Cross-entropy	Pairwise
Luggage	0.020	**0.377**	Shorts	0.042	**0.607**
LongSkirt	0.007	**0.181**	Formal	0.393	**0.444**
Capri	0.011	**0.075**	Blue	0.128	**0.624**
Brown	0.051	**0.366**	Grey	0.565	**0.577**
White	0.121	**0.568**	Jeans	0.258	**0.404**
Plastic Bags	0.018	**0.049**	Black	0.691	**0.745**
Suits	0.312	**0.423**	Casual	0.922	**0.957**
Short Skirt	0.025	**0.269**	Trousers	0.525	**0.658**

Table 5.6: Performance on validation set in terms of the area under curve (AOC) of the precision-recall characteristics for the legs part

Attributes legs	Cross-entropy	Pairwise	Attributes legs	Cross-entropy	Pairwise
Luggage	0.001	0.001	Shorts	0.057	**0.320**
LongSkirt	0.010	**0.057**	Formal	0.018	**0.026**
Capri	0.034	**0.099**	Blue	0.381	**0.573**
Brown	0.029	**0.107**	Grey	0.390	**0.442**
White	0.141	**0.456**	Jeans	0.535	**0.701**
Plastic Bags	0.035	**0.037**	Black	**0.596**	0.587
Suits	0.001	0.001	Casual	0.967	**0.972**
Short Skirt	0.023	**0.275**	Trousers	**0.531**	0.529

Table 5.7: Performance on VIPeR dataset in terms of the area under curve (AOC) of the precision-recall characteristics for the legs part

Attribute	# individuals	Attribute	# individuals	Attribute	# individuals
upperBodyThickStripes	89	lowerBodyBrown	310	upperBodyWhite	1882
hairWhite	99	hairGrey	382	hairLong	2024
carryingLuggageCase	105	accessoryHairBand	394	lowerBodyGrey	2258
upperBodyOrange	106	lowerBodyWhite	398	carryingBackpack	2518
upperBodyYellow	128	carryingPlasticBags	407	lowerBodyJeans	2621
lowerBodyLongSkirt	133	lowerBodySuits	425	upperBodyBlack	3661
hairBald	140	lowerBodyShortSkirt	466	lowerBodyBlack	4293
accessoryMuffler	148	lowerBodyShorts	487	lowerBodyTrousers	4375
upperBodyPink	161	upperBodyBrown	495	hairShort	6598
upperBodySweater	164	upperBodyLogo	497	upperBodyLongSleeve	6785
hairYellow	205	upperBodyJacket	518	hairBlack	6859
lowerBodyCapri	205	upperBodyRed	615	upperBodyCasual	7738
upperBodyThinStripes	222	upperBodyBlue	644	lowerBodyCasual	7786
upperBodyNoSleeve	240	upperBodyTshirt	840		
upperBodyGreen	264	upperBodyFormal	892		
upperBodyPurple	276	lowerBodyFormal	892		
accessorySunglasses	278	hairBrown	1102		
upperBodySuit	305	lowerBodyBlue	1399		
accessoryHat	306	upperBodyShortSleeve	1652		
upperBodyPlaid	308	upperBodyGrey	1709		

Table 5.8: Number of individuals presenting each attribute for the whole PETA dataset

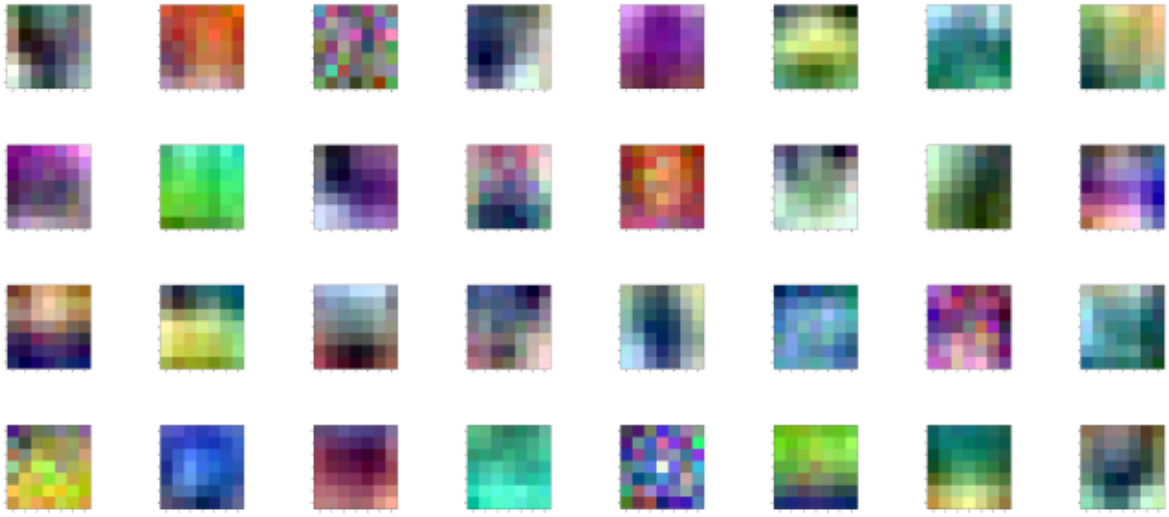

Figure 5.6: The 32 kernels at the first convolutional layer for the network relative to the head part

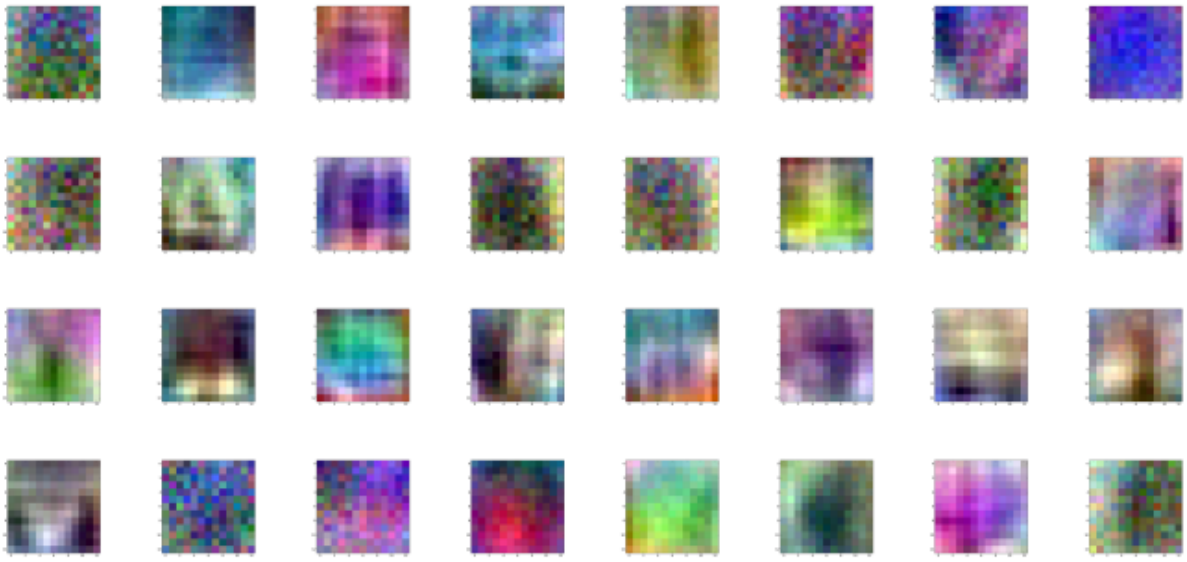

Figure 5.7: The 32 kernels at the first convolutional layer for the network relative to the torso part

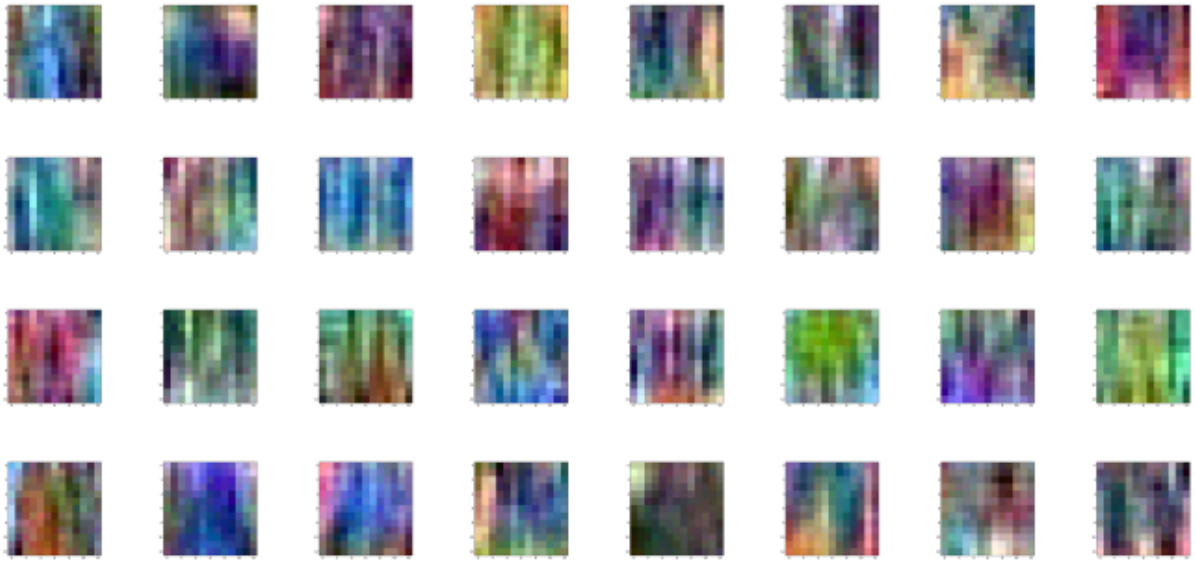

Figure 5.8: The 32 kernels at the first convolutional layer for the network relative to the legs part

5.3 Conclusions and future work

In conclusion, this work consisted in exploiting deep learning techniques for semantic attribute retrieval of pedestrians acquired in video surveillance scenarios. We showed that an objective function that takes into account the co-occurrence of attributes can largely improve the accuracy of detectors, and is also able to generalize to new domains. We point out that there is still room for improvement. First, exploiting techniques that can be used to behave against the highly unbalanced number of positive and negative examples for each attribute. Second, more sophisticated methods for taking advantage of attribute co-occurrence can be used, especially if they are able also to exploit the relations among attributes corresponding to different body parts. Finally, another interesting point is to investigate on more general attributes such as age and sex that has been discarded in this preliminary work.

Figure 5.9: First ten pedestrian returned as response to some significant attribute query, along with the ratio of positive examples in the test set

Figure 5.10: First ten pedestrian returned as response to some significant attribute query, along with the ratio of positive examples in the test set

Lower body jeans (269/1264 examples)

Upper body t-shirt (162/1264 examples)

Upper body thin stripes (10/1264 examples)

Figure 5.11: First ten pedestrian returned as response to some significant attribute query, along with the ratio of positive examples in the test set

Chapter 6

Concluding remarks

In this chapter it will be given an overall conclusion about the work we presented in this dissertation. The main focus has been put to the description of the appearance of an individual, acquired by a video surveillance network. We started considering hand crafted features regarding both the clothing appearance and anthropometric measures taken at a far distance from the pedestrian. Thanks to the Multiple Component Dissimilarity (MCD) framework, we have been able to implement a system for person re-identification using RGB-D cameras, that work in real-time and with state of the art accuracy. However, it is worth giving some words also about the current limitations of the work. First, RGB-D cameras such as Kinect, nowadays are able to work only in indoor scenarios. Second, in terms of the affluence of people in the surveillance zone, such sensors are not able to work in a crowd scene, and because of this, in chapter 3 we considered a limited number of people. Anyway, this is not necessarily a limiting factor in the immediate future. Nowadays, 3D sensor such as Velodine LIDAR[1], are able to capture the 3D in outdoor scenes, in the range of hundreds of meters. At the same time, the version two of the Kinect sensor, is able to track the skeleton of six individuals instead of just two. The framework we proposed is also addressed to these new scenarios, and in the future it will make 3D re-identification possible in more unrestrained settings.

Another interesting point is about the use of more modalities, such as face recognition cues and gait analysis. Especially for face recognition, such algorithms can be applicable only if the pedestrian is giving the front to the surveillance cameras. The use of our MCD framework is limited from this point of view, and a possible future work can consist on making it able to deal with missing modalities. In addition, with the recent breakthrough given by deep learning in image annotation [52], it is possible to encapsulate into MCD also some mid-level features learned automatically from raw data in both unsupervised and supervised way.

With respect to semantic retrieval of pedestrian in surveillance recordings, we can see that in this work, the use of deep learning techniques made it possible to retrieve more complex attributes and generalize to new datasets. As in person re-identification, a possible future work can consist on taking advantage of both hand crafted and mid-level features, in order to take advantage of the potentiality of both approaches. Another point that still has

[1] http://velodynelidar.com/

not being well investigated, is about the learning of pedestrian representations, that consent the re-use of the learned pedestrian model, for instance in person re-identification. Considering that in the worst case, for person re-identification we have for each individual just one frame for training and one frame for test, taking advantage of the knowledge learned from a task with much more data, can potentially boost the performance and robustness of a standalone method. In general, this transfer of knowledge, along with the exploitation of correlation among data, is a hot topic in the computer vision community. Taking apart person re-identification, a system able to well generalize from a small case of study, can be absolutely important for instance in medical application. Therefore, a possible future work can consist on extending the proposed approaches to new datasets, new tasks and fields of research.

Bibliography

[1] A. Albiol, A. Albiol, J. Oliver, and J.M. Mossi. Who is who at different cameras: people re-identification using depth cameras. *Computer Vision, IET*, 6(5):378–387, 2012. [cited at p. 7]

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1014–1021, 2009. [cited at p. 2, 33]

[3] 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. 6, 7]

[4] 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. 6]

[5] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects*, pages 59–64, Scottsdale, Arizona, USA, November 2011. [cited at p. 44]

[6] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. Sarc3d: A new 3d body model for people tracking and re-identification. In Giuseppe Maino and GianLuca Foresti, editors, *Image Analysis and Processing ICIAP 2011*, volume 6978 of *Lecture Notes in Computer Science*, pages 197–206. Springer Berlin Heidelberg, 2011. [cited at p. 44]

[7] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *1st International Workshop on Re-Identification (REID 2012)*, 2012, in press. [cited at p. 7, 8, 18, 19, 21, 23]

[8] 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. 7]

[9] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012. [cited at p. 49]

[10] 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. 7]

[11] C. BenAbdelkader and Y. Yacoob. Statistical estimation of human anthropometry from a single uncalibrated image. *Computational Forensics*, 2008. [cited at p. 7]

[12] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464, June 2011. [cited at p. 44]

[13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. [cited at p. 45]

[14] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation. [cited at p. 49]

[15] Alina Bialkowski, Simon Denman, Patrick Lucey, Sridha Sridharan, and Clinton B Fookes. A database for person re-identification in multi-camera surveillance networks. *Proceedings of the 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA 12)*, pages 1–8, 2012. [cited at p. 9]

[16] 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. *Proc. 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. 7]

[17] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*, 2011. [cited at p. 44]

[18] 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. iv, 5, 6, 7]

[19] Keyang Cheng, Yongzhao Zhan, and Man Qi. Al-ddcnn: a distributed crossing semantic gap learning for person re-identification. *Concurrency and Computation: Practice and Experience*, pages n/a–n/a, 2016. CPE-15-0485.R1. [cited at p. 13]

[20] Chang Chih-chung and Lin Chih-Jen. Libsvm: a library for support vector machines, 2001. [cited at p. 34]

[21] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002. [cited at p. 20]

[22] 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. 7]

[23] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM. [cited at p. 52]

[24] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67. [cited at p. 47]

[25] I.O. de Oliveira and J.L. de Souza Pio. Object reidentification in multiple cameras system. In *4th International Conference on Embedded and Multimedia Computing (EM-Com)*, pages 1–8, 2009. [cited at p. 7]

[26] Yubin DENG, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 789–792, New York, NY, USA, 2014. ACM. [cited at p. v, vii, 44, 45, 49, 51]

[27] 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. 9, 37]

[28] Piotr Dollár, Boris Babenko, Serge Belongie, Pietro Perona, and Zhuowen Tu. Multiple component learning for object detection. *Computer Vision–ECCV 2008*, pages 211–224, 2008. [cited at p. 9]

[29] Husheng Dong, Chunping Liu, Yi Ji, Zhaohui Wang, and Shengrong Gong. Fusion of spatially constrained attributes with kernelized ranking for person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015. [cited at p. 13]

[30] 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. 5]

[31] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367, June 2010. [cited at p. iv, 6, 7, 12, 13, 19, 20, 22, 25]

[32] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of the 2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2360–2367, 2010. [cited at p. 12, 33, 36]

[33] Dario Figueira, Loris Bazzani, Ha Quang Minh, Marco Cristani, Alexandre Bernardino, and Vittorio Murino. Semi-supervised multi-feature learning for person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 111–116. IEEE, 2013. [cited at p. 9, 16, 27]

[34] 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. 7]

[35] Mu Gao, Yuning Du, Haizhou Ai, and Shihong Lao. A hybrid approach to pedestrian clothing color attribute extraction. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 81–84, May 2015. [cited at p. 12, 13]

[36] Amit Garg, Jonathan Noyola, Romil Verma, Ashutosh Saxena, and Aditya Jami. Exploring correlation between labels to improve multi-label classification. *CoRR*, abs/1511.07953, 2015. [cited at p. 13, 47, 48]

[37] 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. 6, 7]

[38] E. Gianaria, N. Balossino, M. Grangetto, and M. Lucenteforte. Gait characterization using dynamic skeleton acquisition. In *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, pages 440–445, Sept 2013. [cited at p. 8, 18]

[39] Elena Gianaria, Marco Grangetto, Maurizio Lucenteforte, and Nello Balossino. Human classification using gait features. In Virginio Cantoni, Dimo Dimov, and Massimo Tistarelli, editors, *Biometric Authentication*, Lecture Notes in Computer Science, pages 16–27. Springer International Publishing, 2014. [cited at p. 8]

[40] 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. 7]

[41] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013. [cited at p. 48]

[42] Doug Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*, 2007. [cited at p. 21, 44, 51, 52]

[43] Douglas Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. of the 10th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 41–47, 2007. [cited at p. iv, 5, 33]

[44] 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. 7]

[45] 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. 27]

[46] 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. 7]

[47] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011. [cited at p. 44]

[48] Home Office Scientific Development Branch. Imagery library for intelligent detection systems (i-LIDS). http://homeoffice.gov.uk, 2007. [cited at p. 21, 44]

[49] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*, volume 2. CRC Press, 2010. [cited at p. 37]

[50] 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. 47, 48]

[51] Ryo Kawai, Yasushi Makihara, Chunsheng Hua, Haruyuki Iwama, and Yasushi Yagi. Person re-identification using view-dependent score-level fusion of gait and color features. In *ICPR*, pages 2694–2697, 2012. [cited at p. 6, 9]

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. [cited at p. 2, 13, 46, 61]

[53] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. *BMVC*, 2(3):8, 2012. [cited at p. 12, 13, 29, 31]

[54] 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. 7]

[55] 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. 7]

[56] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. [cited at p. 12, 13]

[57] L. Li, W. Huang, I.Y.H. Gu, and Q. Tian. Foreground object detection from videos containing complex background. pages 2–10, 2003. cited By 169. [cited at p. 12]

[58] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In KyoungMu Lee, Yasuyuki Matsushita, JamesM. Rehg, and Zhanyi Hu, editors, *Computer Vision ACCV 2012*, volume 7724 of *Lecture Notes in Computer Science*, pages 31–44. Springer Berlin Heidelberg, 2013. [cited at p. 44]

[59] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014, in press. [cited at p. 6]

[60] Chunxiao Liu, Shaogang Gong, and Chen Change Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 47(4):1602 – 1615, 2014. [cited at p. 44]

[61] Javier Lorenzo-Navarro, Modesto Castrillon-Santana, and Daniel Hernandez-Sosa. On the use of simple geometric descriptors provided by rgb-d sensors for re-identification. *Sensors*, 13(7):8222, 2013. [cited at p. 8]

[62] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2648–2655. IEEE, 2013. [cited at p. v, 3, 45, 46]

[63] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2648–2655, 2013. [cited at p. 12]

[64] Bingpeng Ma, Yu Su, and FrÃ©dÃ©ric Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6â[[[ERROR FOR PACKAGE inputenc]]][[[ERROR FOR PACKAGE inputenc]]]7):379 – 390, 2014. [cited at p. iv, 7, 19, 20, 25]

[65] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, page 1, 2013. [cited at p. 49]

[66] C. Madden and M. Piccardi. Height measurement as a session-based biometric for people matching across disjoint camera views. In *Image and Vision Computing New Zealand*, page 29, 2005. [cited at p. 7]

[67] N. Martinel, A. Das, C. Micheloni, and A.K. Roy-Chowdhury. Re-identification in the function space of feature warps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015, in press. [cited at p. 6]

[68] N. Martinel and C. 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, 2012. [cited at p. 6]

[69] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 161–181. Springer London, 2014. [cited at p. 8, 18, 21]

[70] Andreas MÃ¸gelmose, Albert ClapÃ©s, Chris Bahnsen, Thomas B. Moeslund, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *9th IEEE Workshop on Perception Beyond the Visible Spectrum*. IEEE, 2013. [cited at p. 6, 8]

[71] Andreas MÃ¸gelmose, Thomas B. Moeslund, and Kamal Nasrollahi. *Multimodal Person Re-identification Using RGB-D Sensors and a Transient Identification Database*. IEEE, 2013. [cited at p. 6, 9]

[72] 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. 7]

[73] Ngoc-Bao Nguyen, Vu-Hoang Nguyen, Thanh Ngo Duc, Duy-Dinh Le, and Duc Anh Duong. *MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II*, chapter AttRel: An Approach to Person Re-Identification by Exploiting Attribute Relationships, pages 50–60. Springer International Publishing, Cham, 2015. [cited at p. 13]

[74] 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. 27]

[75] 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. 7]

[76] M. Oren, C.P. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *cvpr*, pages 193–99, 1997. [cited at p. 44]

[77] Federico Pala, Riccardo Satta, Giorgio Fumera, and Fabio Roli. Multi-modal person re-identification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 04/2015 2015. [cited at p. 12]

[78] Elzbieta Pekalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005. [cited at p. 2, 10]

[79] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999. [cited at p. 19]

[80] 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. 7]

[81] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. [cited at p. 16]

[82] Riccardo Satta, Giorgio Fumera, and Fabio Roli. Exploiting dissimilarity representations for person re-identification. In *Proceedings of the 1st International Workshop on Similarity-Based Pattern Analysis and Recognition (SIMBAD)*, pages 275–289, 2011. [cited at p. 9, 10, 19, 25, 29, 31, 33]

[83] Riccardo Satta, Giorgio Fumera, and Fabio Roli. Fast person re-identification based on dissimilarity representations. *Pattern Recognition Letters*, 33(14):1838 – 1848, 2012. [cited at p. iv, 2, 7, 9, 10, 11, 20, 25, 26]

[84] Riccardo Satta, Giorgio Fumera, and Fabio Roli. A general method for appearance-based people search based on textual queries. In *First International ECCV Workshop on Re-Identification (ReID 2012)*, Florence, Italy, 12/10/2012 2012. [cited at p. 12]

[85] Riccardo Satta, Giorgio Fumera, Fabio Roli, Marco Cristani, and Vittorio Murino. A multiple component matching framework for person re-identification. In *Proc. of the 16th Int. Conf. on Image Analysis and Processing (ICIAP)*, volume 2, pages 140–149, 2011. [cited at p. iv, 2, 6, 9, 19, 20, 25, 33]

[86] Riccardo Satta, Federico Pala, Giorgio Fumera, and Fabio Roli. Real-time appearance-based person re-identification over multiple kinecttm cameras. In *8th International Conference on Computer Vision Theory and Applications (VISAPP 2013)*, Barcelona, Spain, 21/02/2013 2013. [cited at p. 7]

[87] Riccardo Satta, Federico Pala, Giorgio Fumera, and Fabio Roli. *People search with textual queries about clothing appearance attributes*. Springer, 2014. [cited at p. 12, 13]

[88] A. Schumann and R. Stiefelhagen. Transferring attributes for person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6, Aug 2015. [cited at p. 13]

[89] W.R. Schwartz and L.S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009. [cited at p. 21]

[90] 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. 7]

[91] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014. [cited at p. 47]

[92] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3739–3747, 2015. [cited at p. 13]

[93] 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. 7]

[94] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen. Person attribute search for large-area video surveillance. In *2011 IEEE International Conference on Technologies for Homeland Security (HST)*, pages 55 –61, 2011. [cited at p. 12, 29, 30]

[95] 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. 7]

[96] Daniel Vaquero, Rogerio Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *IEEE Workshop on Applications of Computer Vision (WACV'09)*, 2009. [cited at p. 12, 29, 30]

[97] Jun Wang and Jean-Daniel Zucker. Solving the multiple-instance problem: A lazy learning approach. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1119–1126, 2000. [cited at p. 11, 20]

[98] Liming Wang, Jianbo Shi, Gang Song, and I-fan Shen. Object detection combining recognition and segmentation. In Yasushi Yagi, SingBing Kang, InSo Kweon, and Hongbin Zha, editors, *Computer Vision ACCV 2007*, volume 4843 of *Lecture Notes in Computer Science*, pages 189–199. Springer Berlin Heidelberg, 2007. [cited at p. 45]

[99] Z. Wu, Y. Li, and R. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014, in press. [cited at p. 6]

[100] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. Specific person retrieval via incomplete text description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 547–550. ACM, 2015. [cited at p. 13]

[101] Zhi Zhou, Yue Wang, and Eam Khwang Teoh. A framework for semantic people description in multi-camera surveillance systems. *Image and Vision Computing*, 46:29 – 46, 2016. [cited at p. 12, 13]

List of Publications Related to the Thesis

Le presenti pubblicazioni e questa tesi sono state prodotte durante la frequenza del corso di dottorato in ingegneria elettrica ed elettronica dell'Universita' degli Studi di Cagliari, a.a. 2014/2015 - XXVIII ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitivita' regionale e occupazione, Asse IV Capitale umano, Linea di Attivita' l.3.1 Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell'ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali.

Federico Pala gratefully acknowledges Sardinia Regional Government for the financial support of her PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

Published papers

Journal papers

Federico Pala, Riccardo Satta, Giorgio Fumera and Fabio Roli, *Multi-modal Person Re-Identification Using RGB-D Cameras* in IEEE Transactions on Circuits and Systems for Video Technology, 2015

Conference papers

Riccardo Satta, Federico Pala, Giorgio Fumera and Fabio Roli, *Real-time Appearance-based Person Re-identification Over Multiple Kinect Cameras* in International Conference on Computer Vision Theory and Applications VISAPP, 2013

Book chapters

Riccardo Satta, Federico Pala, Giorgio Fumera and Fabio Roli, *People search with textual queries about clothing appearance attributes* in Person Re-Identification, pp. 371-389, Springer London, 2014