3D body models for people Re-Identification

Davide Baltieri

Department of Engineering

University of Modena and Reggio Emilia

A thesis submitted for the degree of

Doctor of Philosophy in

Computer Engineering and Science

Sunlight bright upon my pillow
Lighter than an eiderdown
Will she let the weeping willow
Wind his branches round

Julia dream, dreamboat queen, queen of all my dreams

Every night I turn the light out
Waiting for the velvet bride
Will the scaly armadillo
Find me where I'm hiding

Julia dream, dreamboat queen, queen of all my dreams

Will the misty master break me
Will the key unlock my mind
Will the following footsteps catch me
Am I really dying

Julia dream, dreamboat queen, queen of all my dreams

Pink Floyd - Julia Dream

Acknowledgements

As i was writing these Acknowledgments i was startled by the number of people that have made all of this a reality, and i have began to appreciate how lucky i was in knowing them and in having the opportunity of calling them *friends*. First i should start with my Ph.D. tutor Roberto Vezzani, who had to endure a lousy student as myself for three years. I should also thank my colleagues at ImageLab for having made my life there bearable: Michele Fornaciari, who has been the greatest friend and companion in many drunken nights tearing down Pomposa, Daniele Borghesani who entertained all of us with his political and weird discussions, Marco Manfredi who had to bear my loud music (even when i was wearing headphones), Simone Calderara for teaching me about life and machine learning, Costantino Grana who probably teached us more about C/C++ than five years of Uni, Dalia Coppi, who was so tired of all of us that escaped to England, Paolo Santinelli, Andrea Prati and obviously our boss, Rita Cucchiara who made all of this possible. Outside of academia i really need to thank my oldest friend and almost brother Stefano Guerra, my flatmates Marco Zanoni, Alessandro Valgimigli and Khan Zhang companions in lots of domestic misadventures, Stephen Donohoe the drunkest irishman ever known to man, and lots of other friends i can't add here for lack of space. And finally the obvious one, my family: my father Mosé and my mother Mirella for enduring having a son like me (and the little fact that they are the sole reason i actually exist). Oh, and my sister Mara (otherwise she will get mad at me for not having thanked her here).

Abstract

One of the most challenging problems in automated surveillance in large camera networks is how to integrate information coming from different cameras; in particular the most recent topic to generate interest in surveillance is **People Re-identification** in images and video archives. People re-identification can be defined as the task of assigning the same identifier to all the instances of the same object or, more specifically, of the same person, by means of the visual aspects that have been captured and extracted from multiple images or videos from different cameras.

In this thesis two novel approaches to people re-identification are presented: SARC3D, based on non-articulated 3D body models, and its extension to articulated 3D body models. The adoption of 3D body models is quite new for re-identification and has not been explored much as a solution in the past. 3D body models allow to spatially map and locate appearance descriptors on a 3D surface. People matching and view integration are directly handled on the 3D body model, reducing the effects of occlusions, partial views or pose changes which normally afflicts holistic 2D descriptors.

Additionally, in the following chapters a thoroughly examination of the state of the art of re-identification algorithms and datasets is given and a novel approach to people orientation estimation is presented, a fundamental step for the correct alignment of 3D models and 2D surveillance images.

A complete experimental evaluation of the proposed methods is reported.

Contents

List of Figures

List of Tables

List of Symbols

When not differently specified, superscript indicates a model identifier, subscript a particular element of a model (i.e. a specific feature vector descriptor, or a specific vertex).

I_t	The current frame	60
B_t	The background model	61
F_t	The foreground image of the current frame	61
x	Position	61
$\mathbf{z}$	A set of observation on the current frame	61
$p(\mathbf{x}\|\mathbf{z})$	Probability density function associated to $\mathbf{x}$	62
$E(\mathbf{x}_t)$	Estimation function, gives the most likely position$\mathbf{x}$	62
w	Weight value	62
$\widehat{N_{eff}}$	Degeneracy values of the particles in a particle filter	62
AM_t	Appearance model used by the particle filter tracker	64
Fx	Fixed part of the appearance model AM_t	64
A_t	Adaptive part of the appearance model AM_t	64
D_t	Dynamic part of the appearance model D_t	64
$\Phi(\cdot)$	Bhattacharyya distance	65
$\vec{v}$	Velocity vector	65
$\mathcal{N}(\cdot)$	Gaussian distribution	65
$\boldsymbol{C}$	Set of discrete orientation	69
c_i	i-th orientation class	69
ψ	Classifier response	70
$\bar{c}$	Predicted discrete orientation	70
$\bar{\theta_k}$	Predicted continuous orientation	70
Ψ	Trained classifier	71
ϕ	A single HoG vector	71
Q	A non overlapping block of the appearance image	71
$\mathcal{WN}(\cdot)$	Wrapped Normal distribution	72
$\mathcal{AWN}(\cdot)$	Approximated Wrapped Normal distribution	73
$MoA\mathcal{WN}(\cdot)$	Mixture of Approximated Wrapped Normal distributions	73
$D_H(\cdot)$	Distance between two SARC3D models	73
$d(\cdot)$	Distance between two feature vector Ξ	74
$d_{Hg}(\cdot)$	Distance between HoG histograms	75
ς	Saliency measure	75
D_ς	Saliency-based distance	75
$D_{H\varsigma}$	Combined feature vector distance with the saliency-based one	76

$\mathcal{T}$	Set of joints	78
τ	Single joint of a joint set $\mathcal{T}$	78
$\mathcal{B}$	Bone set	78
β	Single bone of the bone set $\mathcal{B}$	78
N_β	Number of bones	78
$\mathcal{W}$	Point cloud	78
κ	RGB Color	78
$\mathcal{P}$	Function returning the index of the closest bone	78
α	Bone weight	80
$\mathcal{X}$	PCA of the bone feature vector $\mathbf{H}$	81
§	Reduced bone feature vector	81
d_M	Mahalanobis distance	81
S	Bone split	82

Chapter 1

Introduction

Traditional CCTV networks are today quite diffuse and pervasive in our everyday life; especially since 9/11 security issues have become one of the top priorities for governments around the world, so much so that most city-wide CCTV networks of today have grown to enormous size, and are composed by thousands of cameras or even more. Famous are the cases of London and New York, in which the numbers of active CCTV cameras has reached the millionth mark. Even smaller towns, like Modena, feature more than a thousand CCTV cameras.

In traditional CCTV systems, video streams are transmitted to a central location, displayed on one or several video monitors and recorded. Security personnel observe the video to determine if there is ongoing activity that warrants a response. Given that such events may occur infrequently, detection of salient events requires focused observation by the user for extended periods of time.

However, the volume of information generated by large, modern, CCTV networks is so large, that humans alone cannot process anymore the data they generate. Computer vision offers some of the tools necessary to offload most of the surveillance tasks onto computers; as an example latest commercially available video surveillance systems attempt to reduce the burden on the user by employing video motion detectors to detect changes in a given scene, or to detect trespassing into secured areas.

In the last decade the computer vision research field has witnessed impressive advancements in pattern recognition and machine learning techniques. These advancements have resulted in the production of more effective systems and appli-

cations for both surveillance and forensics industries, and consequentially an ever increasing demand for these products. Systems and tools for forensic analysis of faces, fingerprints and other biometric parameters along with smart surveillance of people and urban environments are spreading; illustrating their relevance to the security industry. This expanding market is a strong catalyst for new research to solve unresolved problems concerning video and multimedia data. Most of the challenges stem from the need to understand the content of an enormous amount of visual data.

We can generally speak of *people analysis* in terms of its relevance to *security*, which has embraced many topics deeply rooted in the field's research over the last decade. Some of these topics include: moving target detection; which has been dealt with in surveillance using background subtraction techniques [168]; people tracking; which exploits time coherency to follow the same object/person along time and space [212]; and people detection by appearance; which adopts machine learning techniques for object (and in particular human and pedestrian) detection [61, 97, 196]. Finally, many studies regarding action and behavior classification that aim to recognize the posture (static), action (short term) or global behavior (long term) of monitored individuals are drawing great interest; as highlighted by Gorelick et al. [93].

One of the most challenging problems in automated surveillance in large camera networks is how to integrate information coming from different cameras; in particular the most recent topic to generate interest in surveillance is **People Re-identification** in images and video archives.

1.1 People Re-Identification

People re-identification can be defined as the task of assigning the same identifier to all the instances of the same object or, more specifically, of the same person, by means of the visual aspects that have been captured and extracted from multiple images or videos from different cameras.

With this premise, people re-identification aims to answer questions such as "Where have I seen this person before?" [215], or "Where has he gone after being caught on this surveillance camera?". In order to understand the role of

people analysis for security, let us first define the terms "detect, classify, identify, recognize, and verify" as provided by the European Commission in EUROSUR-2011 [85] for surveillance:

- *Detect*: to establish the presence of an object and its geographic location, but not necessarily its nature.
- *Classify*: to establish the type (class) of object (car, van, trailer, cargo ship, tanker, fishing boat).
- *Identify*: to establish the unique identity of the object (name, number), as a rule without prior knowledge.
- *Recognize*: to establish that a detected object is a specific pre-defined unique object.
- *Verify*: Given prior knowledge on the object, can its presence/position be confirmed.

In agreement with the EUROSUR definition, re-identification lies in between identification and recognition. It can be embraced in the *identification* task assuming that the goal of re-identification is to match people's aspects using an un-supervised strategy without prior knowledge. One application is the collection of flow statistics and extraction of long-term people trajectories in large-area surveillance. Re-identification allows a coherent identification of people acquired by different cameras and different points of view, merging together the short-term outputs of each single camera tracking system.

Re-identification can also be associated with the *recognition* task whenever a specific query with a target person is provided and all the corresponding instances are searched in a large database. Multimedia forensics searching for a suspect within the database of videos from a crime associated neighborhood or a visual query made in an employer database are practical examples of its application as a soft-biometric tool. This is suitable in cases of low resolution images with non-collaborative targets and when biometric recognition is not feasible.

Re-identification works on the exterior appearance usually acquired by noisy cameras, which makes impossible to extract and associate precise measurements

such as biometric features. For this reason, re-identification methods have to address various hard challenges associated with illumination variability, different video hardware, pose variations, different viewpoint and even changes in the clothing appearance. Figure 1.1 shows some examples of the challenges re-identification methods are faced with: all images shows the same person viewed by different cameras, under different light condition over several days.

Figure 1.1: Different snapshot of the same pedestrian viewed by different cameras, under different light condition over several days

Thus, people re-identification by visual aspect is emerging as a very interesting field and future solutions could be exploited as a tool for soft-biometric technology, long term surveillance, or support for searching in security-related databases.

1.2 Similarities and dissimilarities between Re-Identification, Tracking and Biometric Recognition

Most methodologies of people re-identification are shared with two other well-known approaches; people tracking and biometric recognition. These both require matching multiple instances of the same person in a video sequence but are characterized by different aims. People tracking (and, more generally, object

tracking) mainly focuses on "maintaining an accurate representation of the object state and position given measurements" [84]. On the contrary, biometry is devoted to find the exact identity of each piece of evidence. Different conditions and hypotheses on the spatio-temporal continuity allow us to recognize specific differences between the three themes. If the frame rate is sufficiently high, image patches containing people from consecutive frames of a video sequence will satisfy four different continuity conditions, with exception to small variations in:

- *Position*, both in the 3D space and in the 2D camera image plane;
- *Point of view*, even if the camera is moving;
- *Appearance*, mainly in reference to clothing style, texture and color;
- *Biometric profile*, which is constant and discriminative for each person.

Commonly, tracking algorithms are based on all the previous hypothesis of constancy and they try to solve additional challenges such as illumination changes, noise, occlusions and so on.

Different from people tracking, the re-identification task aims to match people instances during a time delay and/or a change in point of view. These invalidates the first two continuity constraints, while the global appearance, in addition to the biometric profile, are preserved (See Table 1.1).

Thus, re-identification becomes a suitable approach for providing data association when different images of people are captured without a sufficient temporal or spatial continuity. This works best in a scenario with a relatively short time period, guaranteeing the constraint of a similar visual appearance. In reality, re-identification cannot be applied to find similarities among people after several days due to likely alterations in their visual appearance, i.e. a change of attire. Biometric recognition can overcome these constraints by working on highly discriminative and stable features computed on the face, iris and fingerprint.

The distinction between tracking, re-identification and biometry are slowly fading, leaving behind a plethora of methods which fall in-between the two classes. Examples include soft-biometry [109] and tracking algorithms by data association. Some tracking algorithms designed to handle occlusion issues relax temporal

Table 1.1: Continuity constraints imposed by people tracking, re-identification and biometric recognition

	People tracking	People re-identification	Biometric recognition
Continuity of:			
Position	✓	✘	✘
Point of view	✓	✘	✘
Appearance	✓	✓	✘
Biometric profile	✓	✓	✓

continuity constraints and resemble re-identification in the way that they share similar methodologies. In addition, the recent "tracking-by-detection" approaches [9] that aims to link the detections of the same individual without requiring the prediction steps of position or appearance, alleviate the restriction of the first two continuity constraints.

1.3 Thesis Overview

In this thesis a novel 3D based re-identification method and its evolution are presented. Chapter 2 presents a detailed survey of the state of the art in **People Re-identification**, the chapter gives a detailed description of the characterizing aspects and main issues concerning the methods specifically designed for People Re-identification, categorized through a novel multidimensional taxonomy. Chapter 3 presents the datasets, the evaluation metrics and benchmarks for people re-identification algorithms. It also details two new datasets developed by the author of this thesis, namely **the ViSOR dataset** and **the 3DPeS dataset**, two novel datasets specifically designed for people-reidentification algorithms testing. Chapter 4 presents in great details the main proposal of this thesis, a novel re-identification method based on *non-articulated 3D human body models* called **SARC3D**, all the necessary steps leading to the model creation and usage in re-identification are given, together with a thoroughly experimental evaluation on the newly developed dataset **3DPeS**. The new system has been extended to *articulated 3D human body models* as detailed in Chapter 5. A full experimental

evaluation of the proposed methods is reported in Chapter 6.

Chapter 2

Re-Identification: State of The Art

This chapter gives a detailed description of the state of the art, the characterizing aspects and main issues concerning the methods specifically designed for **People Re-identification**. A conceptualization of the re-identification task is provided, by describing in detail the different dimensions of the current problems and previously proposed solutions through a novel multidimensional taxonomy All the issues and challenging aspects of people re-identification are tackled, describing the solutions proposed in the past, starting from the first attempts with holistic descriptors to the more recent 2d body model based approaches.

2.1 Re-identification: a multidimensional overview

Research in surveillance and people analysis for security has been thoroughly focused on people re-identification during the last decade which has seen the exploitation of many paradigms and approaches of pattern recognition. Despite best efforts, no consistent or conclusive results have been published.

To better understand the similarities and commonalities of the approaches at hand, a multidimensional taxonomy of the problem is exploited. Instead of adopting a hierarchical taxonomy where a classification criteria is placed at each level (such as in the work by Aggarwal and Cai [1]), Re-Identification approaches

are categorized through a multidimensional space as illustrated in Figure 2.1. Re-identification approaches can be characterized by differences in Camera Settings, the Sample Set cardinality, the Signature (or feature set), the adoption of a Body Model, the exploitation of Machine Learning techniques, and the Application Scenario.

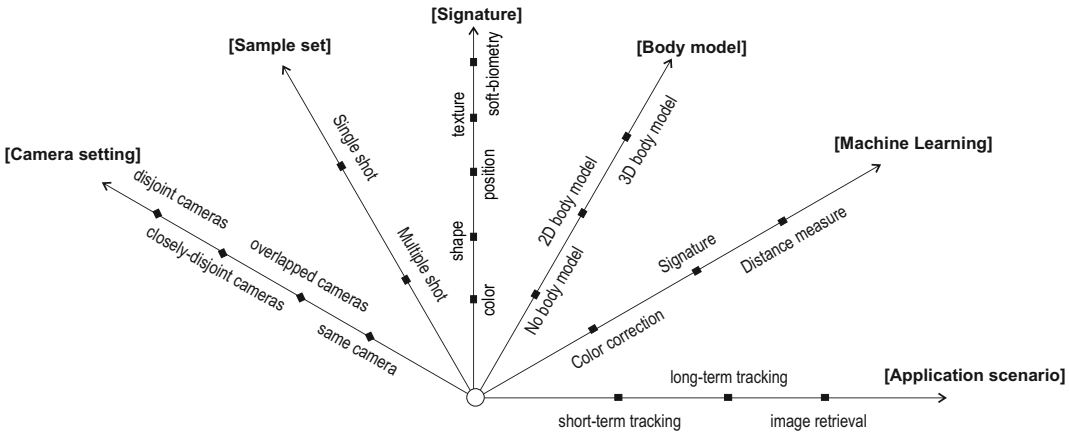


Figure 2.1: Multidimensional taxonomy for people re-identification algorithms

The first relevant dimension is the **Camera Setting**, which defines the type of recorded visual data and the global layout of the available cameras being exploited. The capabilities of a re-identification solution depend on the assumptions made about the known fields of view (FoVs) and the acquisition system. Holding information about the camera setting means the re-identification task can exploit many geometric and temporal relations while also examining color and spatial constraints from different views. We can distinguish four main situations that usually arise: *same camera*, *overlapping cameras*, *calibrated disjoint cameras* and *uncalibrated disjoint cameras*. The last case includes contexts without any knowledge of the camera placement or device setting and incorporates all possible datasets acquired from private image collections, mobile devices, web providers, social networks or any other possible data source.

The second dimension is the cardinality of the **Sample Set**. Depending on the application, scenario and the data availability; the re-identification process can amalgamate multiple samples of the same person. However, the most frequent

cases present us with a single shot of the targeted individual. This situation is typical of current forensics applications of people recognition where the visual aspect is extracted from a single picture or video frame. In the context of video surveillance, the input data produced is a video with known camera settings using a tracking system capable of capturing more images of the same person. The increased availability of shots means it can be utilized as an effective tool in video surveillance. In this dimension, re-identification solutions can be accordingly divided into *Single Shot* and *Multiple Shot* approaches.

One of the most important space dimensions is the **Signature**, which specifies the set of features collected from the samples and used to provide a discriminative profile for each person. Re-identification algorithms are required to extract a compact and representative signature of each detected instance. The feature composing the people signature is one of the most distinctive aspects of all pattern recognition problems. A signature can be based on a single or a combination of features that include *Color*, *Shape*, *Position*, *Texture*, and *Soft-biometry*; all specific to the human form.

Partially related to the previous dimension, the fourth dimension is the spatial level mapping on a **Body Model**. Extracted features can be computed on different spatial levels, varying from holistic features computed on and describing a Region of Interest, to local descriptors extracted from local patches. In the last case, the local descriptors can be grouped in a global set or mapped to a body model, preserving the spatial location of each descriptor. Prior knowledge of the generic human shape and structure can be exploited to localize the extracted visual features. Model-based localization provides a more coherent and accurate representation of the image and grants the correct comparison of corresponding body parts. Problems that arise from occlusions and segmentation errors can be minimized. For instance the straightforward ambiguity of "white shirt and black pants" versus "black shirt and white pants" can be solved. While simplified 2D models have traditionally been the most commonly used, recent years have seen the introduction of few proposals using paradigms based on 3D body models.

More precise body models and signatures usually call for more accurate input data (Region of Interest, RoI). This usually comes from a simple *Bounding box* (BB) obtained with a people detector, to a pixel-wise *Silhouette* (SIL) obtained

by a precise foreground segmentation, to a set of *Body parts* (BP) segmented and classified from the person silhouette in a deterministic, fixed or learned manner (see examples in Fig. 2.2).

The fifth dimension corresponds to the exploitation of **Machine Learning** algorithms during the re-identification process. In the past, this approach has been used in three different steps; learning the color transformation among different cameras, creating a more discriminative signature and tuning the distance metric among samples. Supervised or semi-supervised Machine Learning algorithms greatly improve re-identification performance, but require a fully representative training set.

The specific **Application Scenario** is the case where the re-identification task is exploited and does happen to define some constraints and peculiarities. For example, in outdoor surveillance the image resolution is usually insufficient for the computation of biometric features or the adoption of a detailed 3D body model. At the same time, real-time processing in surveillance may guarantee sufficient frame synchronization for the management of multiple overlapped cameras.

Fig. 2.1 summarizes the six dimensions which provide us with over a thousand different combinations of parameters and solutions. However, not all the combinations are significant and feasible. In the following sections we provide a review of the literature that is consistently classified with the proposed six dimensional space.

Figure 2.2: From left to right: Region of Interest: bounding box by people detection, complete pixel-wise silhouette from foreground segmentation, face/body part segmentation and classification. Examples of body models: three horizontal fixed slices [5, 123, 160], ten slices [34], symmetry based parts [80], five human body parts [19]

2.2 Fifteen years of research in re-identification

Since the paper by Cai and Aggarwal [41] first described an attempt to follow the same person using a multi camera scenario, roughly one hundred papers on the subject have been proposed, which have been classified and summarized (see Tab. 2.1). Instead of providing a detailed description of each proposed method, our discussion focuses on the peculiarities, requirements, and advantages of the various techniques using the reference multidimensional taxonomy introduced in the previous section.

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
RE-IDENTIFICATION FOR TRACKING					
[121]	Same Med-res. Indoor/Outdoor	Single	Color [RGB], Shape, Texture *HOG and Covariance*	none	Caviar, Trecvid08 and ETH
[39]	Same Med-res.	Single	Color [RGB], Shape *RGB Histogram, Gradients and Optical Flow*	none	ETHZ Central, TUD crossing, i-Lids, UBC Hockey, ETHZ Soccer
[122]	Same Calib. Disj. Med-res.	Multiple	Color [RGB], Shape, Texture *RGB Histogram, Covariance Matrix, HOG*	none	CAVIAR, TRECVID08
[188]	Calib. Disj. Med-res.	Single	Shape *RGB Histogram*	none	CAVIAR, VideoWeb
[90]	Calib. Disj. Med-res. Indoor	Single	Color [RGB] *RGB Histogram*	none	p.v.
[30]	Overlapping Med-res.	Single	Color [RGB] *RGB Histogram*	none	p.v.
[198]	Same Indoor/Outdoor	Single	Color [n.a.], Texture *Head appearance*	none	p.v.
[209]	Same Indoor	Single	Color [rg] *Color histogram, face and voice*	none	p.v.

continued on next page

continued from previous page

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[65]	Same Uncal. Disj. Calib. Disj. Hi-res.	Single	Color [HSV] *bag of soft biometric features: Eye, hair, skin and clothes color, eyeglasses, beard and mustache presence, weight*	none	p.v. , Caviar, ViPER
[4]	Uncal. Disj. Hi-res. Indoor	Multiple	Color [RGB] *Bodyprints*	3D - 90 stripes	p.v.
[144]	Uncal. Disj. Med-res. Indoor	Single	Color [RGB,YCbCr,HSV], Texture *Color, Schmid and Gabor texture features*	none	i-LIDS (MCTS)
[141]	Overlapping Med-res. Outdoor	Multiple	Color [HSV], Position *Color histogram and feet position*	none	PETS2009, Caviar
[56]	Uncal. Disj. Med-res. Outdoor	Multiple	Color [RGB] *Appearance mask*	none	PETS2010
[114]	Uncal. Disj. Med-res. Indoor	Multiple	Texture *SIFT and ISM*	none	CasiaA
[115]	Same Med-res. Indoor	Multiple	Texture *SIFT on infrared images*	none	Casia Infrared dataset
[3]	Calib. Disj. Lo-res. Indoor/Outdoor	Single	Color [several], Shape, Texture, Position	2D - grids	Viper, p.v.
[132]	Overlapping Hi-res. Outdoor	Single	Position *Adaptive homographies at different levels*	none	p.v.
[55]	Uncal. Disj. Hi-res. Indoor	Multiple	Color [RGB] *Color-position histogram*	2D - LTH with fixed slices	p.v.
[203]	Calib. Disj. Lo-res. Outdoor	Single	Color [RGB], Position *Color histogram and position*	none	ViSOR
[12]	Overlapping Hi-res. Indoor	Single	Color [n.a.], Position *Position on the ground plane and image appearance*	none	Sport videos
[129]	Overlapping Hi-res. Indoor	Single	Position *Vertical axis and homography projections*	none	p.v.
[206]	Overlapping Lo-res. Outdoor	Single	Color [RGB] *Color histogram*	2D - body parts	p.v.

continued on next page

continued from previous page

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[112]	Overlapping Lo-res. Outdoor	Single	Position *Feet position*	none	PETS2001
[95]	Uncal. Disj. Lo-res. Outdoor	Single	Color [RGB, HSV, YCbCr], Texture *Mean values*	rectangular stripes	ViPER
[105]	Same Med-res. Indoor	Single	Color [RGB], Texture *SIFT and color autocorrelogram*	none	Caviar
[108]	Overlapping Med-res. Outdoor	Single	Position *Feet position and body height*	2D rectangular with fixed height	p.v.
[111]	Calib. Disj. Hi-res. Outdoor	Multiple	Color [RGB], Position *Feet position and brightness transfer function*	none	Online cameras
[44]	Overlapping Lo-res. Outdoor	Single	Shape, Position *Feet and head position, vertical axis*	none	ViSOR
[52]	Uncal. Disj. Med-res. Indoor	Single	Color [YCbCr], Texture *Mean color, cov. matrix and others*	none	Torino metro station
[49]	Uncal. Disj. Calib. Disj. Med-res.	Single	Color [RGB] *RGB histogram + calibrazione + camera network topology*	none	p.v.
[140]	Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB] *MCSHR: color clusters*	none	p.v.
[214]	Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB], Shape *Color path-length profile*	none	Honeywell
[220]	Overlapping Med-res. Outdoor	Single	Position *Feet position*	none	PETS2001
[136]	Calib. Disj. Outdoor	Single	Shape, Position *Width, height, motion and position*	none	p.v.
[106]	Overlapping Med-res. Outdoor	Single	Position *Principal axis*	none	PETS2001, p.v.
[86]	Uncal. Disj. Med-res. Indoor	Multiple	Color [n.a.] *Appearance Map*	3D - Cylinder - PAM	p.v.

continued on next page

continued from previous page

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[165]	Uncal. Disj. Med-res. Indoor	Single	Color [RGB] *Color histogram*	none	p.v.
[5]	Uncal. Disj. Med-res. Indoor	Single	Color [n.a.], Texture *Spatial and spectral distribution of dominant colours*	2D - LTH	Online cameras
[215]	Same Med-res. Indoor	Single	Color [RGB] *Mean color*	2D - LTH with fixed slices	Icra05
[118]	Overlapping Med-res. Outdoor	Single	Position *Feet position*	none	PETS2001, p.v.
[35]	Overlapping Calib. Disj. Outdoor	Multiple	Position *Feet position*	none	PETS2001, p.v.
[47]	Overlapping Lo-res. Indoor	Single	Color [HSV], Position *Feet and head position, color GMM, height*	none	Online cameras
[120]	Same Med-res. Indoor	Multiple	Color [RGB] *Color histogram for each region of the scene*	none	Online cameras
[117]	Calib. Disj. Lo-res. Indoor	Single	Color [HSV], Position *Colour, speed and spatio-temporal camera model*	none	p.v.
[43]	Overlapping Lo-res. Indoor	Single	Position *Feet position and mean intensity*	none	Online cameras
[157]	Same Uncal. Disj. Med-res. Outdoor	Single	Color [YUV] *Color histogram*	none	p.v.
	RE-IDENTIFICATION FOR RETRIEVAL				
[29]	Uncal. Disj. Med-res. Indoor/Outdoor	Multiple	Color [HSV], Texture *Histogram Plus Epitome*	2D - LTH plus simmetry based vertical splits	i-LIDS, ETHZ, CAVIAR4REID
[134]	Uncal. Disj. Med-res. Indoor/Outdoor	Single	Color [RGB, YCbCr,HSV], Texture *Color, Schmid and Gabor texture features*	2D - 6 stripes	i-LIDS (MCTS), ViPER
[124]	Uncal. Disj. Hi-res. Indoor/Outdoor	Single	Color [] *High level attributes*	none	i-LIDS (MCTS), ViPER, ETZH

continued on next page

continued from previous page

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[104]	Uncal. Disj. Med-res. Indoor/Outdoor	Single	Color [HSV, Lab], Texture *Mean color and LBP histogram*	2D - Grid	ViPER, ETZH, Prid2011
[64]	Uncal. Disj. Hi-res. Indoor	Single	Color [HSV], Texture *Texture and color*	none	Feret
[23]	Uncal. Disj. Med-res. Indoor/Outdoor	Multiple	Color [HSV] *Color histogram*	3D model	ViSOR, Sarc3D
[207]	Uncal. Disj. Hi-res. Indoor	Single	Color [RGB] *Color histogram*	2D - upper and lower part	p.v.
[18]	Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB], Shape *MRCG - Mean Riemaniann Covariance Grid*	none	i-LIDS, ETHZ
[218]	Uncal. Disj. Med-res. Indoor/Outdoor	Single	Color [RGB, YCbCr,HSV], Texture *Color, Schmid and Gabor texture features*	2D - 6 stripes	i-LIDS (MCTS), ViPER
[27]	Uncal. Disj. Lo-res. Indoor	Multiple	Texture *SIFT, SURF, SC, GLOH*	none	Caviar
[83]	Same Hi-res. Indoor	Single	Texture *DCT-based facial appearance*	2D - face	p.v.
[15]	Uncal. Disj. Hi-res. Indoor	Single	Texture *SIFT, SURF, Spin*	2D - LTH	p.v.
[80]	Same Uncal. Disj. Indoor/Outdoor	Multiple	Color [HSV], Shape, Texture *Weighted color histograms, MSCR, recurrent high structured patches*	2D - LTH plus simmetry based vertical splits	ViPER, i-LIDS, ETHZ
[40]	Uncal. Disj. Med-res. Indoor/Outdoor	Single	Color [RGB, HS,YCbCr], Texture *Color, Schmid and Gabor texture features*	2D - 6 stripes	i-LIDS (MCTS), ViPER
[19]	Uncal. Disj. Med-res. Indoor	Single	Color [RGB], Texture *Haar based and DCD based signature*	2D - body parts	i-LIDS (MCTS)
[146]	Uncal. Disj. Lo-res. Outdoor	Single	Color [Several], Shape *Color Histogram*	none	ViPER

continued on next page

continued from previous page

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[6]	Uncal. Disj. Lo-res.	Single	Color [RGB, YCbCr,HSV], Texture *RGB, YCbCr,HSV histogram, Schmid and Gabor filters responses*	none	ViPER
[217]	Uncal. Disj. Hi-res. Indoor	Single	Color [RGB], Texture *CRRRO descriptor: SIFT and color*	none	i-LIDS (MCTS)
[150]	Overlapping Uncal. Disj. Lo-res. Indoor	Single	Color [CIEluv] *Mean Color*	2D - LTH with fixed slices	Online cameras
[66]	Uncal. Disj. Lo-res. Indoor	Single	Color [modified HSV], Texture *SURF*	none	Caviar, Weizmann
[135]	Uncal. Disj. Med-res. Indoor	Multiple	Color [YCbCr], Texture *Bag of siftch - color SIFT*	none	Caviar, ViPER
[133]	Uncal. Disj. Calib. Disj. Med-res. Indoor	Multiple	Color [rgb] *Color rank*	none	Honeywell dataset
[101]	Uncal. Disj. Lo-res. Indoor	Multiple	Texture *Set of SURF-like descriptors*	none	Caviar
[183]	Uncal. Disj. Indoor/Outdoor	Single	Color [YCbCr], Texture *SIFT and MPEG7 color layout*	none	TRECVID
[167]	Uncal. Disj. Med-res. Outdoor	Single	Color [RGB] *Weighted color histogram*	2D - standard human mask	Online cameras
[88]	Uncal. Disj. Med-res. Outdoor	Single	Color [modified HSV, RGB], Texture *Appearance Map*	2D - spatio temporal appearance model	p.v.
[160]	Overlapping Uncal. Disj. Calib. Disj. Med-res. Outdoor	Single	Color [HSV], Shape, Position *Mean color, feet position, height, bodybuild ratios*	2D - LTH with fixed slices	p.v.
[34]	Same Med-res. Outdoor	Single	Color [HSL] *Median color*	2D - 10 horizontal slices	p.v. - bus stop

continued on next page

continued from previous page

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[123]	Uncal. Disj. Med-res. Indoor	Single	Color [HSV], Texture *Dominant color, histograms, edge energy*	2D - LTH with fixed slices	p.v.
[151]	Uncal. Disj. Lo-res. Indoor	Single	Color [RGB, rgb], Shape *Color histogram and shape features.*	none	p.v.
[219]	Uncal. Disj. Med-res. Indoor/Outdoor	Single	Color [RGB, HS,YCbCr], Texture *Color, Schmid and Gabor texture features*	2D - 6 stripes	i-LIDS (MCTS), ViPER, ETZH
[69]	Same Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB] *Height of the LTH parts and color histograms*	2D - LTH	PETS2006

Table 2.1: Examples of re-identification methods classified with the multidimensional taxonomy (grouped by main application scenario and in chronological order)

2.2.1 Camera Setting

Apart from some initial experiments of people re-identification as a particular case of shape classification (e.g. the seminal work by Cai and Aggarwal [42]), most of the proposals come from surveillance and forensics scenarios where assumptions can be made about the camera settings. Additionally, some algorithms have been previously proposed that automatically reveal the topology of available cameras and thus can recover the setting parameters. For example, Niu and Grimson [154] present a statistical method to learn the environment's topology using a large amount of tracking data;Calderara et al. [44] and Khan and Shah [118] proposed the use of camera hand-offs of people walking to detect and estimate the camera overlapping. For additional details on the automatic discovery of the camera network topology, please refer to the survey by Radke [174].

According to the previous taxonomy, re-identification proposals can be divided on the basis of knowledge or assumptions on camera topology.

Same camera: in this setting, the goal of re-identification is to be able to

identify the same individual repeatedly using the same camera after he/she is initially detected [34, 83]. The main constraints relate to the point of view, which is considered unchanging. The proposed approaches are similar to the ones for disjointed cameras but assume a single acquisition source. This simplifies the matching task since challenges of view point discrepancies and color distortions are neglected. However, this is not a limitation since several applications are based on this specific setting, such as the control of a set of entrance gates or an indoor environment that is monitored by a single camera. The work by Bird et al. [34] describes an application in public transportation areas, while the recent work by Jungling and Arens [115] designs re-identification on the basis of infrared cameras. Finally, re-identification could be very useful as support for single camera tracking where excessive occlusions of extensive time periods frequently occur. In these cases, a model of the scene and the occluding obstacles could improve the matching performances as described by Gong et al. [91].

Overlapping cameras: in this scenario, re-identification can be considered as a part of a long-term tracking process over enlarged fields of view. This problem is also called *consistent labeling* [42, 44, 106, 118]. Geometrical properties and relations among cameras can be exploited after a full or partial calibration of the system. Overlapping cameras operate under the assumption that the detections being matched are captured at the same instant by different cameras. If this fails to occur, the overlapping property comes out to naught. When several cameras are capturing the same region, the re-identification process can even operate in crowded scenarios where multiple individuals are occluding one another [119]. The first work on people matching and re-identification was proposed by Cai and Aggarwal [41] fifteen years ago and assumed the presence of a layout composed by overlapping cameras. The original proposal has been successively improved and formalized [42, 43], and now relies on the geometrical relations and constraints intrinsically embedded in multiple views of the same object or scene [102]. Implementing the epipolar relationship largely utilized in stereo vision, it is possible to reduce the number of potential matches. For people surveillance and forensics, the primary assumption made is that people's feet maintain permanent contact with the ground while walking and thus the position of the feet in each view can be analytically mapped to all other views. More

specifically, the coordinates of a person's feet in different views are related by homography transformations which are defined for each couple of static cameras. . An initial manual or automatic [44, 118] estimation of the homography matrix allows the re-identification problem to be reduced to geometric based matching of the feet coordinates. A plethora of systems based on this assumption and relationship have been proposed; [12, 42, 78, 106, 108, 119, 125, 150].

Calibrated disjoint cameras: this setting is oriented toward large area surveillance with known camera layout. Even if the cameras' fields of view are non- overlapping, some geometrical information can still be useful[3, 35, 111, 117]. Using a homography transformation to obtain feet position on a common ground plane [125], the temporal gap between two corresponding views can be bridged [142]. This is made possible by means of predictive filters such as the Kalman filter [78] or the particle filter [203]. In addition, temporal relations could be used to refine the selection of candidates based on the time gap between captures as proposed by Mazzon et al. [144].

Uncalibrated Disjoint cameras: this is the most general, yet complex case. No assumptions or predictions are made by virtue of the cameras position. They can be installed over a wide range in a multitude of diverse settings and conditions; indoor/outdoor, wide/narrow, field of view, etc. featuring non-homogeneous capabilities and technologies. In this case the re-identification task is sometimes referred to as *re-acquisition*, as suggested by Cong et al. [54]. A large number of proposals have addressed the a-posteriori color calibration and/or transformation of inhomogeneous cameras. The color distribution of a person can vary significantly when captured by different cameras. Section 2.2.4 provides more details on this problem and proposed solutions are included. With no limitations on the camera setting, images and videos can be collected by unknown devices (typical of available web video) or by unconstrained mobile devices. Re-identification by aspect similarity can be considered a form of soft-biometry for people identification which has been a useful person recognition tool for social networking applications. In this case, geometrical information is not available and only the person's appearance can be used as a matching feature [55, 88]. As illustrated in Table 2.1, the majority of works submitted in recent years fall into this category and address re-identification without any setting limitations.

Additional aspects of the camera setting, such as the point-of-view (viewing angle), image resolution (number of pixels), and the image quality (compression, noise, etc.) strongly affect the re-identification framework such that they limit the remaining dimensions of the taxonomy. For instance; a low image resolution could prevent the adoption of soft-biometric features or a top-view camera setting could make the mapping of appearance features on 3D body models more difficult, and so on. More details are reported in the following.

2.2.2 Sample Cardinality

The re-identification efficacy is related to the amount of information available in terms of both image resolution and number of available samples. Single shot methods associate pairs of images only, with each pair containing a single shot of an individual's appearance. Methods of the second class (multiple shots) take advantage of information coming from multiple frames depicting the same person [29]. Single shot techniques are more general and can be applied to a wider range of applications. Conversely, multiple shot algorithms reach a more complete and invariant signature which is potentially more promising. However, multiple shot algorithms lack in the sense that they require additional tasks, and are often computationally severe for both data alignment and dimensionality reduction. While the majority of the algorithms belong to the Single Shot class (e.g.,[5, 19, 123, 160, 167]), information below describes some examples of the multiple shot strategies designed recently to overcome single shot limitations.

Temporal sampling: A number of key-frames are selected from the individual's history, for instance; Cong et al. [55] selected ten key frames. The feature sets computed on each frame are concatenated before a spectral analysis step is applied to reduce the final signature dimensionality. A similar approach has been proposed by Yu et al. [214] that's based on a video key-frame selection.

Set of signatures: If more than one view of the same person is available, a suitable signature is computed and stored for each of them. The classifier works by considering the entire set of available signatures, as suggested by Farenzena et al. [80]. When more than one view of the same person is available at the same time (i.e., the layout is composed by more overlapping cameras), camera

switching and/or best view selection strategies can be used in order to select the most distinguishable view [43]. The best view selection is also effective in the presence of occlusions [119].

Set of specialized signatures: a set of signatures is computed and stored for each person. Differently from the previous case a custom signature is computed for each value of a selected parameter and added to the set. For example, the parameter could be the distance from the camera, the person's orientation or the camera tilt. During the classification step, the value of the parameter is measured or estimated and used to retrieve the coherent signatures. It is possible to store a specific signature for each person's orientation from various angles or positions in space. The method proposed by Krumm et al. [120] computes a different training signature for each (discretized) person's position in the scene and each time considers the subset of appearances extracted from the that position.

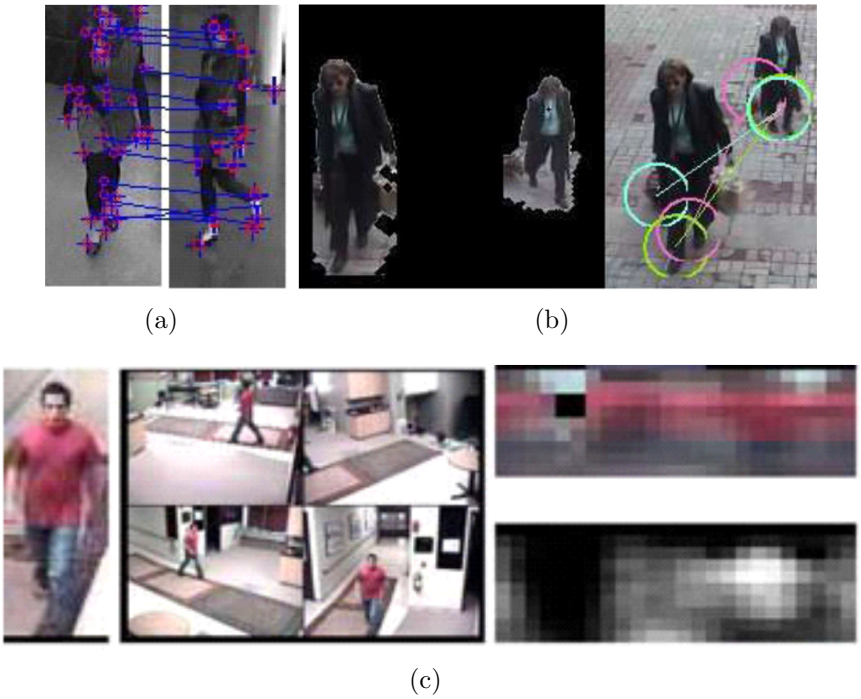
Set of local descriptors: a global set of descriptors computed on local feature points (such as SURF or SIFT) is generated from all available views. The person's signature is defined as the set of descriptors or codebook based histograms (e.g., [101, 115, 135]).

Body-model based signature: the final signature directly integrates more contributions (e.g. [21, 50, 86]). For example, the PAM appearance map developed by Gandhi and Trivedi [86] is obtained by updating the visible part of the signature.

2.2.3 Signature

As with most pattern recognition problems, re-identification efficacy is directly affected by the type of adopted **signature**. Works proposed until now have exploited different features which can be grouped approximately by: (a) color, (b) shape, (c) position, (d) texture, and (e) soft-biometry. Recently, Doretto et al. [75] provided a valuable review of different appearance signatures.

The selection of the adopted feature is determined by different factors. On one side, the signature should be unique or as distinctive as possible which can lead the selection toward biometry or soft-biometry features. On the other side; camera resolution, computational load and other implementation issues can prevent or



(a) (b)

(c)

Figure 2.3: Examples of proposed solutions: (a) SPIN, SURF and SIFT local features [15], (b) Interest operator matching in 2D body models [75], (c) 3D Panoramic map with overlapped cameras [86].

limit their usage and more generic features are required.

Color: Even if it depends on external illumination, camera technology and setting, *color* is the features exploited the most. Plain means, histograms [140, 151, 206, 207] and Gaussian models [52] are some of the possible descriptors on classical color spaces (e.g., RGB, rgb, HSV). The color spaces adopted in numerous studies are illustrated in Table 2.1. In some works, more color descriptors are mixed or compared, such as that found in the works of [199], [95], [40], and [218].

Shape features including width, height, width/height ratio [107], vertical axis [44, 106], moment invariants [148], and contours [213] have been proposed.

Position: The position in the image or on the ground plane is commonly adopted to match people in setups with overlapping cameras [44, 106, 118].

Texture: Covariance matrices [19], SIFT [105, 217] and SURF [101] descriptors are some examples of texture based features (see Fig. 2.3). More recently, HoG like descriptors [156, 184] and LBP features [104] have been proposed.

Mixed descriptors: combination of features have been evaluated with the goal of integrating color, shape, and texture contributions [3, 80, 95, 146] in the same signature. de Oliveira and Pio [66] were able to improve the base SURF descriptors using HSV color information to create an integrated color and texture feature set. Ali et al. [6] computed Schmidt and Gabor texture filters on different color spaces. Kang et al. [116] defined an invariant descriptor that integrates both color and edge contributions. Given a detected moving blob, a reference circle is defined by the smallest circle containing the blob. This circle is uniformly sampled into a set of control points, and for each control point, a set of concentric circles of various radii defines the bins of the appearance model. Inside each bin, a Gaussian color model is computed to model the color properties of the overlapping pixels of the detected blob. The normalized combination of distributions obtained from each control point defines the appearance model of the detected blob. The spatiograms adopted by Birchfield and Rangarajan [33] are a generalization of histograms that includes higher order spatial moments. For example, the second-order spatiogram contains in histogram bin the spatial mean and covariance. A detailed list of adopted features is illustrated in Table 2.1.

Soft-biometry: Iris scanning [149], Palm-Vein images [221], fingerprints, hand appearance [77] and other hard biometric identifiers [67] are expressly excluded from this survey, but intermediate soft-biometric features [109] such as gait [103, 186], facial features [64, 83, 161], body size [69] and body weight [201] are included. These features can be effectively analyzed for re-identification if the image resolution is sufficiently high [63] or if a RGBD sensor is available [26]. Different from hard biometric signatures, soft-biometry lacks the distinctiveness and permanence to identify an individual with high reliability. Soft-biometry depends on physical or behavioral traits typically described as labels and measurements that can be more easily understood. In some cases, soft-biometry allows retrieval and recognition based solely on human descriptions [124, 177, 181].

An overview of the general topic of soft biometry and a new refined definition of the field has been provided by Dantcheva et al. [65], who also propose two novel soft biometric traits, namely based on weight and dress color.

Current research has a particular leaning toward the adoption of "out-of-

the-box" machine learning techniques, which are used to select and integrate simple features into a more complex signature (see Section 2.2.4). Following the aforementioned, recent proposals have shown no desire for a manual feature selection and provide the above mentioned features to the learning module [104]. Despite the fact that re-identification still performs with some level of mediocrity, great improvements could be made with the incorporation of new view invariant descriptors.

2.2.4 Machine Learning

Machine learning techniques can be put into use to automatically discover relations, behaviors and models directly from the data. In re-identification, machine learning algorithms have been adopted on three different levels: at the image level for color correction, at signature level for dimensionality reduction or generating codebook-like descriptors, and at matching level to learn a specific distance measure.

Color correction.

Under the conditions of different camera types or where the illumination is not uniform, people matching using color based signatures can be affected by systematic variations in the input signals (see Fig. 2.4(a)). Several techniques have been proposed to help learn and apply color transformations between different cameras, some using color patterns similar to those reported in Fig. 2.4(b).

A way of matching appearances using different cameras is by finding a transformation that maps colors in one camera to those in the other cameras. With this in mind, linear algebraic models [178] as well as more complex non-linear approaches [89] have been implemented. Despite dependence on a large number of parameters, Javed et al. [111] proved that all such transformations lie in a low dimensional subspace for a given pair of cameras and they propose to estimate the probability that the transformation between current views lies in the learned subspace.

Black et al. [36] used a non-uniform quantization of the HSV color phase to improve illumination invariance, while Bowden and KaewTraKulPong [37] and Gilbert and Bowden [90] exploited the "Consensus - Color Conversion of the

Munsell color space" (CCCM), a coarse quantization based on human perception. Porikli [170] adopted a non linear transformation function for each set of cameras that is learned during a training phase. The transformation is applied to each pixel color or directly to the color histogram bins during matching.

Colombo et al. [52] propose a method for estimating the appropriate transformation between each camera's color space using the covariance of the foreground data collected from each camera; thus applying a second order normalization of both the chromaticity and intensity. Instead of finding a transformation function to be computed, stored and applied for each camera pair, Metternich et al. [146] and van de Sande et al. [199] present us with a set of descriptors and color spaces which appear invariant to the illumination conditions. When illumination changes are solely responsible for the incoherence of the colors among cameras, Brightness Transfer Functions can be an invaluable tool learned and applied initially, as suggested by Porikli [170] and improved by Javed and Shafique [110] and Gilbert and Bowden [90].

Signature computation: An important consideration should be taken on the *role of machine learning* in the signature computation. Recently, sets of local features have taken precedence over holistic or region-wise descriptors, calling for the implementation of feature selection and space dimensionality reduction algorithms. With this in mind, different machine learning based techniques can be applied depending on the computational constraints imposed by the application, which usually require online real-time processing or offline batch learning. Examples can be collected from the tracking field where the exploitation of machine learning is brisker. Kuo et al. [122], Babenko et al. [16], Mei and Ling [145]

(a) (b)

Figure 2.4: A common problem of multi-camera systems: a. different views have different colors; b. Patterns used for the color calibration.

create and update the object model using online learning algorithms which do not require any previous training. Conversely, Grabner et al. [94] and Pellegrini et al. [163] exploit previous knowledge of the object model or surrounding context to improve the tracking reliability. If a batch data process is permissible as in most forensic applications, time consuming algorithms such as CRF models [208] could be applied. Teixeira and Corte-Real [194] introduced an on-line learning step using a bag-of-features model based on SIFT descriptors. Similarly, Babenko et al. [16] proposed creating the appearance model of each person using a Multiple Instance Learning (MIL) algorithm derived from MIL-Boost by Viola et al. [205].The main drawback of these techniques is the requirement of multiple source images required to adopt a specific "class" that corresponds to a specific person in re-identification. Another inconvenience is that off-line computations do not usually permit a fast automatic update mechanism when new examples are provided. Bazzani et al. [29] proposed a novel descriptor for person re-identification that condenses multiple shots into a highly informative signature called the Histogram Plus Epitome, HPE. An image epitome is the result of an image or a set of images collapsing into a small collage of overlapped patches through a generative model that ultimately embed the essence of the textural, shape and appearance properties of the data [113]. A completely different approach has been adopted by Satta et al. [182]. Each individual is represented as a vector of dissimilarity values from a set of learned visual prototypes. Even if the re-identification accuracy is lower than other approaches, particularly when the number of prototypes is low; the trade-off between processing time and accuracy is still advantageous. This presents us with an application for real-time scenarios.

Distance learning: In the past, machine learning has been frequently neglected by re-identification as the majority of reviewed papers seemingly apply common distance metrics and nearest-neighbor approaches to re-identify the same person (e.g., [19, 21, 86]). The focus was traditionally on the actual feature vectors, targeting descriptors as invariant and general as possible. The Bhattacharyya or the Euclidean distance functions are usually adopted depending on the specific feature type. Occasionally, a linear combination with suitable weights has been defined to merge different contributions(e.g., [80].) Recently, more attention has been devoted to learning a good metric. Dikmen et al. [72] proposed a

SVM framework to obtain an optimized metric for nearest neighbor classification called Large Margin Nearest Neighbor (LMNN). Zheng et al. [218] introduced a novel Probabilistic Relative Distance Comparison (PRDC) model, which differs from most existing distance learning methods in that, rather than minimizing intra-class variation whilst maximizing inter-class variation, it aims to maximize the probability of a pair of true match having smaller distance than that of a wrong match pair. An extension of the original method has been presented by the same author in [219]. Kuo et al. [122], Li et al. [131] designed boost learning frameworks to generate an affinity model that is exploited for people's tracklet association. Similarly, Yang et al. [208] handled the association problem using a CRF model. In [104], a distance matrix M is estimated automatically from a training set and then used during the matching steps, similar to the Mahalanobis distance function. Through M, the body parts considered to have the highest priority are selected and assigned higher weights. This approach is called Relaxed Pairwise Metric Learning (RPML) and it has proven to be a highly efficient and effective metric learning approach. RPML aims to compute a pseudo-metric M similar to the Mahalanobis distance; providing a dissimilarity score between two feature vectors. Machine learning approaches have undoubtedly improved re-identification performance in recent years, opening a plethora of new ways to solve surveillance problems. They constitute the most active topic within people re-identification.

2.2.5 Spatial level mapping on a Body Model

People re-identification is a matching problem among "objects" having the same or similar elements of shape and structure. For the most part, appearance based techniques adopt color and texture features more than other geometrical features, which are usually shared by many individuals. At the same time, since body shape can be easily generalized, the adoption of a more simplified body model is normally very effective and useful. A body model can be exploited to spatially map the extracted visual features and thus obtain a more coherent and representative feature set that can be correctly compared.

With an available body model, extracted local descriptors can be mapped

directly to the model while preserving their spatial location within the body (*Mapped local features* [80, 123]). Contrarily, in the absence of a body model; *Global Features* such as global color histograms and shape are the descriptors most often computed and exploited [157]. These holistic features have the advantages of all aggregated measures: reduced sensitivity to noise, low computational cost and no alignment or segmentation steps are required. However, in many instances their ability to discriminate is limited and the specific information embedded in the appearance details cannot be fully exploited.

Hybrid solutions have been proposed which adopt *Unmapped local features* where local descriptors are initially computed on patches or blocks having been collected without preserving any spatial reference (e.g., Bag-of-Words with SIFT descriptors [135]).

Among others, the *cylindrical* shape and the *legs-torso-head* structure are the most widely utilized body model in surveillance and forensics. By modeling a person as a cylindrical shape (or more generally as a solid of revolution), the horizontal variations of a person's appearance can be neglected as color or texture distribution along the vertical axis contains the most significant data. For instance, Bird et al. [34] divided the person's silhouette into ten horizontal stripes with the mean color of each stripe being stored as representative feature.

The reason for the legs-torso-head model is primarily due to traditional western style clothing. The targeted silhouette is divided into three horizontal parts, which ideally correspond to legs (and thus to the pants/skirt appearance), torso (i.e., shirt or jacket) and head (i.e., hair). This segmentation can be accomplished using fixed sizes [5, 91, 123, 150, 160]. Albu et al. [5] placed the cuts at 30% and 80% of the total height, while Monari et al. [150] opted to make cuts at 15% and 70% mark respectively. Other methods did not divide the RoI into fixed parts but propose alternative solutions. Farenzena et al. [80] automatically computed the cut points from profile histograms and split the torso and legs into two parts using symmetry based algorithm. Cheng et al. [50] adopted the Pictorial Structure technique proposed by Andriluka et al. [10] where parts are localized, and their descriptors are extracted and matched. When multiple images of an individual are available, they proposed an algorithm (Custom Pictorial Structure - CPS) to customize the fitting of the pictorial structures on a specific person. Finally, Bak

et al. [19] adopted a body part detector to extract the position of the head, torso and each limb. This case requires a high quality data source and a body part detector involving extensive computation. Figure 2.2 provides an illustration of examples mentioned above. The body models are also reported.

Over the last few years, 3D models have been largely neglected for the re-identification task. One of the few attempts to incorporate a 3D body model was done so by Gandhi and Trivedi [86], where a cylindrical surface (Panoramic Appearance Map) maps local color descriptors. Since then, different graphical models have been reviewed in the literature for 3D people tracking, motion capture, and posture analysis [11, 53]. These models are frequently complex, requiring fine fitting techniques in order to obtain a perfect match between a 3D model posture and the real model.

2.2.6 Application scenarios

The last dimension of the taxonomy regards the application scenarios, which mainly belong to two different areas: surveillance and forensics.

Long-term Tracking. Time constraints are one of the main issues in automatic video surveillance, which differs from the basic CCTV recording by the off-hand detection of events and alarms. In this context, the re-identification task has been used for *long-term tracking*: people should be tracked as long as possible, using one or more cameras. Thanks to the short (or null) temporal gap between samples, geometric and positional features are usually enough and the requirements on the camera quality and resolution are loose. Video sequences are usually available as input and processed using the common surveillance chain composed by background/foreground segmentation and intra-camera object tracking. Data is collected and merged in a "late-fusion" like manor rather than being integrated before the tracking. Consistent labeling approaches proposed by Khan and Shah [118], Calderara et al. [44] and more recently Lian et al. [132] and Madrigal and Hayet [141] belong to this category, as all methods require at least a partial overlapping among the camera fields of view.

Image retrieval. In forensics, the real-time constraint is no longer problematic since the computation is bordering being off-line and human interaction

is permissible. Given a query item, all the frames/images corresponding to the same person should be retrieved. The re-identification task is thus employed for *image retrieval* and usually provides ranking lists, similarly related items, and so on. Complex features and heavy learning algorithms are employed to generate a model for the query item and for the database of eligible candidates. Increasing complexity of both features and matching algorithms enumerates additional constraints on the cameras quality, image resolution and zoom ability. Color based approaches like that of [34, 146] can be applied to any resolution. The method of Farenzena et al. [80] requires medium resolution to find symmetry axes and texture patterns while the work of Fischer et al. [83] based on soft-biometry usually requires more defined source images to capture details and perform precise measurements.

Short-term tracking. Finally, the more recently proposed "tracking-by-detection" approach [9] tries to link the provided detections of the same person by means of re-identification algorithms. In this case, the re-identification task works in terms of *short-term tracking*, similar to a "pure" re-identification approach that requires a feature based signature for each detection. Breitenstein et al. [38] compared a set of different color (RGI/RGB/HS/Lab) and texture (LBP/Haar) features while Brendel et al. [39] adopted a PCA projected vector of HOG descriptors and HSV color histograms. The detections are then connected by means of data association algorithms such as the greedy Hungarian algorithm [164], network flow [216], or spectral clustering [57]. However, the view or temporal gap assumed by the definition of re-identification (see chapter 1) should be null or at least limited. This application scenario is very close to tracking and so the following section and Table 2.1 illustrates some specific examples. We also refer the reader to specific surveys on the topic, such as the work by Yilmaz et al. [212].

2.3 Re-Identification in Today's Research

The goal of this section is to analyze the direction research has tended toward in the last few years. As highlighted in Fig. 2.5, many proposed methods share common templates. Firstly, the camera setting affects the choice of signature and

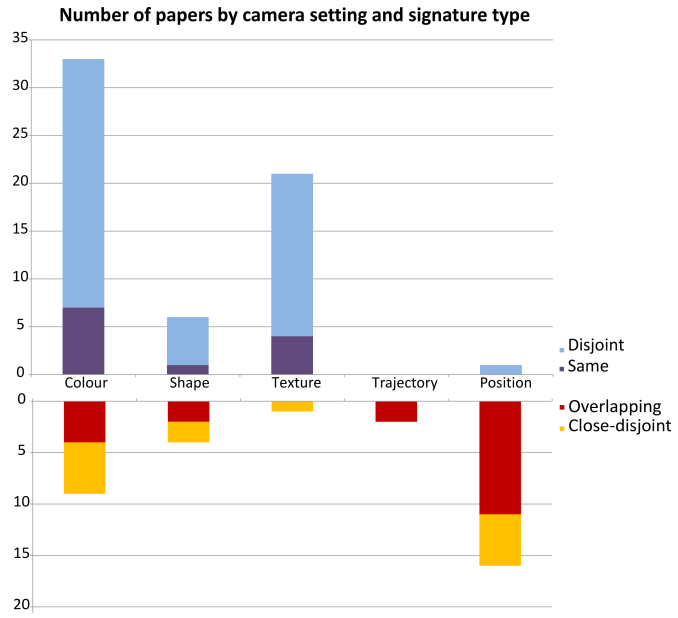


Figure 2.5: Histogram of the features used in the reviewed approaches

the methods of similarity assessment. The geometric position is most reliable if the camera's fields of view are overlapped or partially overlapped (see lower histogram in Fig. 2.5). Instead, for disjoint cameras or unknown single cameras, view-dependent appearance features based on color and texture are adopted more than shape and size.

If the computational resources are limited or the image resolution is low (as in most of the current surveillance videos), holistic features are most often adopted. Holistic features such as color histograms were initially proposed by Javed and Shafique [110]. Some improvements have been made more recently, an example being the introduction of more sophisticated matching criteria between histograms and color correction functions for compensating differences between cameras and views [140, 173].

Only recently have more complex signatures based on texture been proposed and with very promising results, more so with SIFT [19, 194] or SURF [101] descriptors (see Fig. 2.3(a)). They do require a moderately high image resolution and the same resolution requirement holds with methods handled by a body model.

The most promising solutions postulated in recent years based on human mod-

els have been proposed by Farenzena et al. [80],Doretto et al. [75], and Gandhi and Trivedi [86]. Farenzena et al. [80] takes advantage of human vertical symmetry to subdivide the appearance images into five regions. For each region, three features describing complementary aspects of the human appearance are extracted: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent textures. A nearest neighbor matching schema is then applied.

The work by Doretto et al. [75] reviews several appearance descriptors and proposes a part-based signature which in testing outperformed holistic descriptors.

The dependence on people orientation is the main drawback of 2D models (i.e., the rotation angle with respect to the vertical axis) since the reasonable assumption of standing postures permits the division of the human model into horizontal segments only. The cylindrical model based Panoramic Appearance Map proposed by Gandhi and Trivedi [86] allow a fully view-invariant appearance description. As previously stated, these methods are generally more time consuming, requiring an evaluation of the orientation of the people moving within the space. New solutions in this area (see for instance the paper by Odobez and Bouthemy [155]) should be very useful to improve both accuracy and speed.

Over the last few years, the role of machine learning has become a central component in the computer vision field. People re-identification is in align with this trend and current research is primarily devoted to the learning aspects of this area. As described in section 2.2.4, specific feature sets like SDALF [80] have been replaced by a set of standard color and texture descriptors [6, 40, 134, 144, 218, 219]. The selection of important features are then processed by machine learning algorithm in an explicit [134] or implicit way [130, 219].

In addition to machine learning, the diffusion of high resolution cameras and low cost range sensors like the Microsoft Kinect have paved new avenues to the surveillance and forensics research fields, including that of re-identification. Limitations mentioned previously that past researches have failed to overcome can now be eradicated. Articulated motion capture addressed in the past on color images [71, 169, 180] and, more recently, the use depth streams [185, 193] are nowadays feasible in real time. In exploiting this information, new methods for

re-identification will be available. Soft- biometry features can be extracted from the tracked skeleton stream and used to generate a person signature. For example, Barbosa et al. [26] defined a set of ratios of joint distances as a person signature. Albiol et al. [4] took advantage of the body model to generate a color histogram for each vertical stripe. Differently to the methods presented in the past [34] and described in section 2.2.5, the stripes relate to the real body model and not to the captured image, which depends on the camera point of view. However, the current noise level on the estimation of joint position do not allow yet acceptable performance. These methods still require extensive research.

Finally, the recent work of Layne et al. [124] proposing an attribute based approach is worthy of mention. Taking inspiration from the operating procedures of human experts [200], they moved the re-identification from low-level features to medium or high level attributes. It would now be closer to the human description yet more difficult to define in an unambiguous way. This attribute based signature can be also used when a description is provided as a verbal identikit.

Chapter 3

Datasets and Evaluation Metrics

Evaluation is a foundational problem in research. We should capitalize on the lessons learned by decades of studies in computer architecture performance evaluation, where different benchmarks are designed, such as *benchmark suites* of real programs, *kernel benchmarks* for distinct feature testing and *synthetic benchmarks*. Similarly, in computer vision and multimedia, benchmark datasets are defined to test the efficacy and efficiency of code and algorithms. The purposes are manifold.

Although such kernel benchmark exists for famous and assessed problems in computer vision, such as tracking, shadows detection or face recognition, to date kernel benchmarks designed specifically for re-identification are not available.

This chapter includes an exhaustive treatise of the available datasets (Sec. 3.1) and evaluation metrics (Sec. 3.4) for the benchmarking of people re-identification algorithms.

Two new publicly available datasets developed by the author and specifically designed for testing re-identification methods are presented: the **ViSOR** dataset (Sec. 3.2) and the **3DPeS** dataset (Sec. 3.3).

3.1 Datasets

While several datasets are publicly available for testing camera tracking, action classification systems, or for surveillance (see a short review by Vezzani and Cuc-

chiara [202]) and multimedia [158]; no dataset has been designed specifically for testing and benchmarking of re-identification algorithms, and few can actually be adopted for re-identification evaluation, especially for multiple shot techniques or 3D body models.

The following paragraphs highlight some of the datasets that have been adopted in the literature analyzed in the previous chapter (see chapter 2.2).

ViPER Currently, one of the most popular and challenging datasets to test people re-identification as image retrieval is ViPER [96]; which contains 632 pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions (see Fig. 3.1(a)). The data set was collected in an academic setting over the course of several months and each image is scaled to 128x48 pixels. Due to its complexity and the low resolution images, only a few researchers have published their quantitative findings on ViPER. In actuality, some matches are hard to identify by a human, an example being the third couple in Fig. 3.1(a). Currently, the best results on this dataset have been obtained by Farenzena et al. [80] on a subset of the dataset and Gray and Tao [95] who are the dataset's original authors. VIPER can't be fully employed for evaluating methods exploiting multiple shots, video frames, or 3D models since only pairs of bounding boxes of the same person have been collected. The performance of several proposals in reference to this dataset is summarized in section 6.1.

I-LIDS. The I-LIDS Multiple-Camera Tracking Scenario (MCTS) [153] was captured inside a busy airport arrival hall. With an average of 4 images for each person, it contains a total of 476 shots of 119 people captured by multiple non-overlapping cameras. Many of these images undergo large illumination changes and are subject to occlusions. The I-LIDS dataset has been exploited by Bak et al. [18], Brosser et al. [40], and Zheng et al. [218] for a performance evaluation of their proposal.

CAVIAR4REID. This is a small dataset specifically designed for evaluating person re-identification algorithms by Bazzani et al. [28]. It derives from the original CAVIAR dataset, which was initially created to evaluate people tracking and detection algorithms. A total of 72 pedestrians (50 of them with two camera views and the remaining 22 with one camera only) are captured in a shopping center scenario. The ground truth has been used to extract the bounding box

(a)

(b)

Figure 3.1: Shot examples from (a) ViPER [96] and (b) ETZH [184]. ViPER contains a couple of cropped images for each person, while ETZH is composed by full frames (left) and the bounding box annotation to crop the person images (right).

of each pedestrian. For each pedestrian, a set of images from each camera view (where available) is provided in order to maximize the variance with respect to changes in resolution, light, occlusions, and body position; so as to maximize the challenge for re-identification.

ETHZ. The ETHZ dataset for appearance-based modeling was generated by Schwartz and Davis [184] from the original ETHZ video dataset [79]. The original ETHZ dataset was used for human detection and is composed of four video sequences. Samples of testing sequence frames are shown in Fig.3.1(b). The ETHZ dataset presents the additional challenge of being captured by moving cameras. This camera setup provides a range of variations in people's appearances, with significant changes in pose and illumination.

TRECVid 2008. In 2008, the TRECVid competition released a dataset for Surveillance applications captured inside an airport. Roughly 100 hours of video surveillance data was collected by the UK Home Office at the London Gatwick International Airport (10 days * 2 hours/day * 5 cameras). Approximately 44

individuals were detected and matched through the 5 cameras.

PETS2009. The dataset presented by the 2009 edition of the International Workshop on Performance Evaluation of Tracking and Surveillance has been acquired by a multi-camera system and contains sequences with different crowd activities in a real-world environment. Each sequence involves a subset of eight available cameras and up to approximately forty actors.

Videoweb Activities Dataset [68]. The Videoweb Activities Dataset is composed of roughly 2.5 hours of video footage taken by multi-camera systems in realistic scenarios and contains people performing numerous repetitive and non-repetitive tasks. Data was collected over four days using a subset of 37 outdoor wireless cameras from the VideoWeb camera network. Each day is represented by a varying number of scenes containing actions performed by multiple individuals.

ISSIA Soccer dataset:. While not specifically designed for re-identification purposes, this dataset presents us with six synchronized views acquired by six Full- HD cameras during a soccer match [74]. The high similarity among players of the same team makes the intra-view tracking and the re-identification tasks very challenging.

Other datasets proposed by single authors not available for public access have not been referenced in this section and they are rather mentioned in the last column of Table 2.1. For additional references to surveillance datasets, please refer to [202] or the Cantata Project repository [172].

The main benchmarks for re-identification are summarized in Table 3.1.

Name & Ref	Image/Video	People	Additional info
ViPER [96]	Still Images	632	Scenario: Outdoor Place: Outdoor surveillance People Size: 128x48 vision.soe.ucsc.edu
I-LIDS [153]	Video [fps=25] 5 cameras PAL	1000	Scenario: Outdoor/Indoor Place: Collection from different scenarios People Size: 21x53 to 176x326 www.ilids.co.uk
I-LIDS-MA [153]	Still Images PAL	40	Scenario: Indoor Place: Airport People Size: 21x53 to 176x326 www.ilids.co.uk
I-LIDS-AA [153]	Still Images PAL	119	Scenario: Indoor Place: Airport People Size: 21x53 to 176x326

			www.ilids.co.uk
CAVIAR4REID [28]	Still Images 384x288	72	Scenario: Indoor Place: Shopping centre People Size: 17x39 to 72x144 www.lorisbazzani.info
ETHZ [184]	Video [fps=15] 1 cameras 640x480	146	Scenario: Outdoor Place: Moving cameras on city street People Size: 13x30 to 158x432 http://homepages.dcc.ufmg.br/~william/
ViSOR dataset [23]	Still Images 704x576	50	Scenario: Outdoor Place: University Campus People Size: 54x187 to 149x306 www.openvisor.org
3DPeS dataset [24]	Video [fps=15] 8 cameras 704x576	200	Scenario: Outdoor Place: University Campus People Size: 31x100 to 176x267 www.openvisor.org
TRECVid 2008 [187]	Video [fps=25] 5 cameras PAL	300	Scenario: Indoor Place: Gatwick International Airport - London People Size: 21x53 to 176x326 www-nlpir.nist.gov/projects/tv2008/
PETS2009 [166]	Video [fps=7] 8 cameras 768x576	40	Scenario: Outdoor Place: Outdoor surveillance People Size: 26x67 to 57x112 www.cvg.rdg.ac.uk/PETS2009/
Videoweb Activities Dataset [68]	Video [fps=30] 8 cameras 640x480	16	Scenario: Outdoor Place: Courtyard and intersections People Size: 32x62 to 86x170 www.ee.ucr.edu/~amitrc
ISSIA Soccer dataset [74]	Video [fps=25] 6 cameras 1920x1080	25	Scenario: Outdoor Place: Soccer match People Size: 42x82 to 57x130 www.issia.cnr.it/soccerdataset.html

Table 3.1: Datasets available for people Re-identification

3.2 The ViSOR dataset

The interest of the research community in creating reference datasets for performance analysis is always very high. Although new datasets, collecting large amounts of video footage are spreading in surveillance and forensics, few benchmarks with annotation data are available for testing specific tasks like re-identification and 3D/multi-view analysis. For this reason the ViSOR dataset[1] was introduced,

[1]The ViSOR dataset is available here: http://www.openvisor.org/sarc3d.asp

as a means of testing multiple-shot/multiple-view re-identification methods. The dataset contains shots of 50 people and consists of short video clips captured with a calibrated camera. To simplify the model-image alignment; four frames for each clip corresponding to predefined positions and postures of the people that were manually selected. The annotated data set is composed by four views for each person, 200 snapshots in total. Additionally, a reference silhouette is provided for each frame (some examples are shown in Fig. 3.2). The dataset was firstly introduced in [23]. Although useful for initial testing of multi-view methods, the presence of only four fixed views and only 50 different peoples limits the benchmarking capabilities of the dataset.

Figure 3.2: Sample silhouettes from the ViSOR re-identification dataset [23]

3.3 The 3DPeS dataset

Given the limitations of the datasets previously mentioned the author recently developed and released to the public, the 3DPeS[1] dataset[24]. 3DPeS was developed specifically for testing re-identification algorithms, with the aim of overcoming the

[1]The 3DPeS dataset is available here: http://www.openvisor.org/3dpes.asp

limitations and constraints of the available datasets. 3DPeS provides a large volume of data that, in addition to people re-identifications, allows to tests all the usual steps in video surveillance; segmentation, tracking, etc.

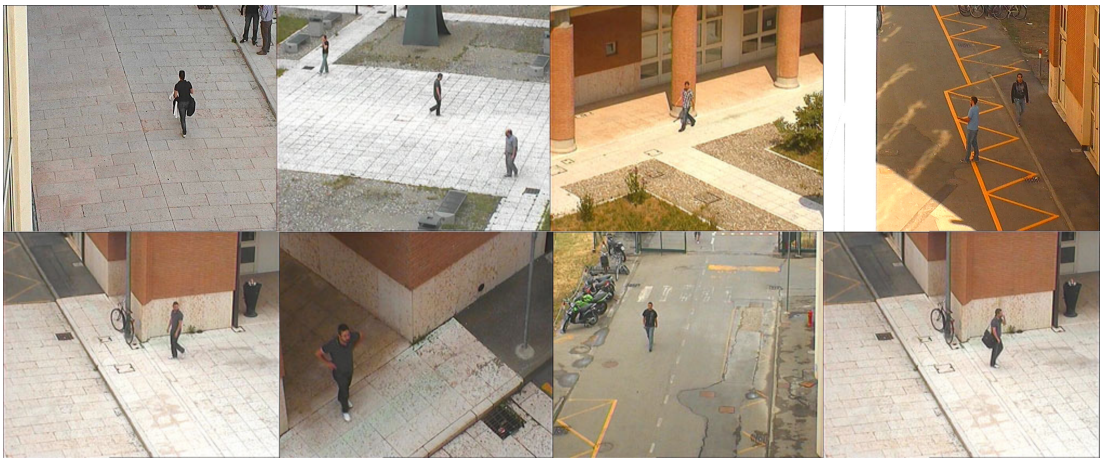


Figure 3.3: Sample frames from 3DPeS [24]

The dataset is captured by a real surveillance setup and is composed of 8 different surveillance cameras (Fig.3.3) monitoring an area of the University of Modena and Reggio Emilia's (UNIMORE) campus. Data was collected over the course of several days. The illumination between cameras is almost constant, but people were recorded multiple times during the course of the day, in clear light and in shadowy areas, resulting in strong variations of light conditions in some cases. The quality of the camera hardware is in line with current standards in visual surveillance, all cameras were from the same vendor and are partially calibrated (position, orientation, pixel aspect ratio and focal length are provided for each one of them). The quality of the images is mostly constant, uncompressed images with a resolution of 704x576 pixels. Depending on the camera position and orientation, people were recorded at different zoom levels. Multiple sequences for 200 individuals are available, together with reference background images, the person bounding box at key frames and the reference silhouettes for more than 150 people.

Table 3.2 reports some quantitative characteristics of the dataset. Annotation comprises: camera parameters, person IDs and correspondences across the dataset, bounding box of the target person in the first frame of the sequence,

Figure 3.4: Sample images from the 3DPeS dataset

preselected snapshot of people appearances (see Fig. 3.4), silhouette for each person snapshot, orientation and bounding box for each person snapshot and a coarse 3D reconstruction of the surveilled area (Fig. 3.5). Each video sequence contains only the target person or a very limited number of people.

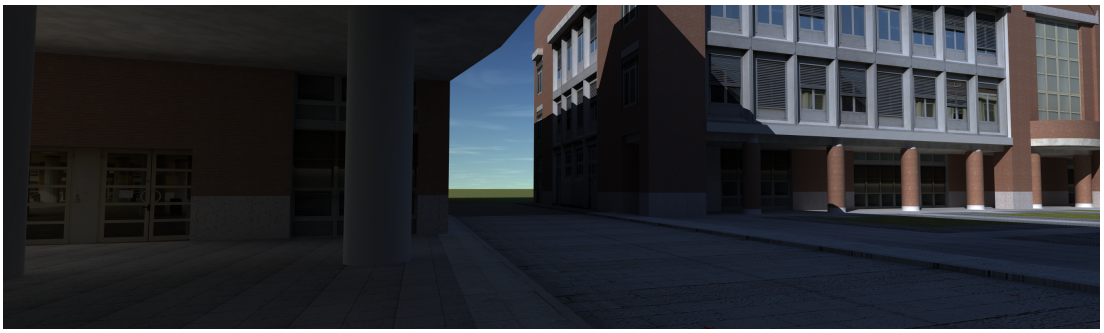


Figure 3.5: 3D Reconstruction of 3DPeS Surveilled Area

Nf videos	612
Nf frames (average, per video)	291
Nf peoples	200
Total Nf frames	178429
Resolution	704x576
Nf video sequences (average, per person)	3
Nf cameras (average, per person)	2

Table 3.2: Quantitative characteristics of the 3DPeS Dataset

3.4 Metrics for Performance Evaluation

In addition to the selection of the testing data, performance evaluation requires suitable metrics depending on the specific goal of the application. According to the definitions introduced in Section 1.1 [85], different metrics are available that relate the specific implementation of re-identification as identification or recognition.

3.4.1 Re-identification as Identification

Since the goal is finding all the correspondences among the set of people instances without *prior* knowledge, the problem resembles data clustering. Each expected cluster is related to one individual. Differently from content-based retrieval problems, where there are relatively few clusters and very large amount of data for each cluster, here the number of desired clusters is very high with respect to the number of elements in each one. However, the same metrics adopted for clustering evaluation could potentially be introduced [8]. *Purity* is one of the widest accepted metrics and is computed by taking the weighted average of maximal precision values for each class. It penalizes the noise in a cluster in instances where one person is wrongly assigned to another individual, but it does not reward grouping together different items from the same category. *Inverse Purity* focuses on the cluster with maximum recall for each category. In other words, it aims to verify all the instances where the same person is matched together and

correctly re-identified.

The performance evaluation of re-identification algorithms is usually simplified, taking into account a group of items at any given moment. The system should state if two items belong to the same person (similarly to the verification problem). In this case, *Precision* and *Recall* metrics applied to the number of hit or miss matches have been adopted [101].

Tasks of re-identification in long-term tracking also fall in this category, especially in surveillance with a network of overlapped or disjoint cameras. With a tracking system, the re-identification algorithm should generate tracks for as long as possible, avoiding errors such as identity switch, erroneous split and merge of tracks, over and under segmentation of traces. For detection and tracking purposes, the ETISEO project [152] proposed some metrics that could potentially be adopted in re-identification. ETISEO was a project devoted to performance evaluation for video surveillance systems, studying the dependency between algorithms and the video characteristics. Sophisticated scores such as the *Tracking Time* and the *Object ID Persistence* have been proposed. The first one corresponds to the percentage of time during which reference data is detected and tracked. This metric gives us a global overview of the performance of the multi-camera tracking algorithm but a problem exists where the evaluation results depend not only on the re-identification algorithms but on the detection and single camera tracking. The second metric regards the re-identification precision, evaluating how many identities have been assigned to the same real person.

Finally, let us cite the work of Leung et al. [127] about performance evaluation of re-acquisition methods specifically conceived for public transport surveillance. Their method takes into account prior knowledge of the scene and normal people behavior in an attempt to estimate how the re-identification system can reduce the entropy of the surveillance framework.

3.4.2 Re-identification as Recognition

In this category the re-identification task aims to provide a set of ranked items given a query target, with the main hypothesis being one and only one item of the gallery can correspond to the query. This is typical of problems faced

by forensics analysts during an investigation where large datasets of image and video footage must be evaluated. The overall re-identification process could be considered a ranking problem [96] where the *Cumulative Matching Characteristic* (CMC) curve is the proper performance evaluation metric, showing how performance improves as the number of resulting images increases. The CMC curve represents the expectation of finding the correct match in the top n matches. From the CMC another common performance measure can be extracted, the AUC (Area Under Curve), defined as the area under the CMC curve. Since the adopted definition of re-identification as recognition given by [85] recalls the definition of identification for biometrics, the evaluation metrics defined in biometrics could be taken into account. Two biometric elements are associated with the same source if their similarity score exceeds a given threshold. Accordingly, the measures of false-acceptance rate (FAR), false-rejection rate (FRR) [115], and the decision-error trade-off (DET) curve can be evaluated, whether or not two snapshots are associated to the same person.

3.4.3 Re-identification in forensics

The precision/recall, FAR, FRR, and DET metrics are now standard and widely accepted in the academic and industrial setting for biometrics and content-base image retrieval. They are yet to be accepted by the legal system in which the court incidentally encounters them on a regular basis. While image analysis is widely adopted during an investigation, final legal judgment comes down to the traditional use of an expert's verbal decision.

Great efforts are being made to improve this practice by adding an objective, quantitative measure of evidential value [147]. With this aim, a likelihood ratio has been suggested for solving different forensics problems like speaker identification [92], DNA analysis [20], and face recognition [7].

The likelihood ratio is the ratio of two probabilities of the same event with different hypotheses. For events A and B, the probability of A given that B is true, divided by the probability of event A given that B is false gives a likelihood ratio. In forensic biology, for instance, likelihood ratios are usually constructed with the numerator being the probability of the evidence if the identified person

is supposed to be the source of the evidence itself, and the denominator being the probability of the evidence if an unidentified person is supposed to be the source. Similar discussions were introduced in a survey for face recognition in forensics [7].

Chapter 4

Non-Articulated 3D Body Models for People Re-Identification: SARC3D

In this chapter a novel 3D based approach for **People Re-identification** is presented. It is designed for typical real surveillance settings, where multiple cameras, often with disjoint fields of view (FoV), can catch the presence and the appearance of many people walking and crossing a monitored environment. This method was previously reported in [21] and [23].

4.1 Introduction: Why 3D Body Models?

People Re-identification is a fundamental task for the analysis of long-term activities and behaviors of specific people. It is becoming one of the major challenges in visual surveillance and forensics due to its intrinsic difficult: the appearance of a person can vary a lot across a distributed network of surveillance cameras because of widely varying camera viewpoints and orientations, illumination conditions, human poses, rapid changes in part of the clothes appearance, occlusions and more: the co-occurrence of many people, crowded scenes, the presence of artifacts or internal and external furniture which occlude the view, the different unideal field of views and the unpredictable motion of people. What can be surely

acquired with the state of the art of computer vision algorithms is, just in some frames, the silhouette of pedestrians, by means of people detectors [98, 196] followed by segmentation [31, 211] algorithms or with some background suppression based [58, 60] people trackers. In this context, a new, robust approach based on people appearance and 3D geometry is proposed.

The adoption of 3D body models is quite new for re-identification and has not been explored much as a solution to this problem, differently from other computer vision fields, such as motion capture and posture estimation [11, 53]. Three basic assumptions are given:

- The frames are acquired by cameras with similar color response (thanks to color calibration, see section 2.2.4, or similar hardware) and the different views are taken under similar illuminants. Since the matching is based on colors, if the cameras behave very differently in color response or have an insufficient color resolution the accuracy will degrade noticeably.

- The people shape, more or less precisely segmented from its background, is available. In this manner specific aspect features of the person appearance can be extracted. The approach accepts possible under-segmentation, which commonly arises with background subtraction methods, and over-segmentation (such as shadows) which are normally neglected.

- Some geometric information about the camera setting, as a simple camera calibration, is available to set the parameter of the 3D model correctly (namely position, height of the model and its pitch and roll angles, as highlighted by fig. 4.1).

The first assumption is implicitly accepted by all re-identification algorithms: since people shapes are actually very similar each other, color is the one most discriminant features and it must be assumed to be somewhat reliable. The same person with the same clothes acquired by the same point of view by two cameras should have the same color. This assumption is ideal since normally cameras have different color responses and people can be filmed under very different light conditions during the course of a day or multiple days, thus colors must be constrained to remain similar and so as to be sufficiently discriminant. The different

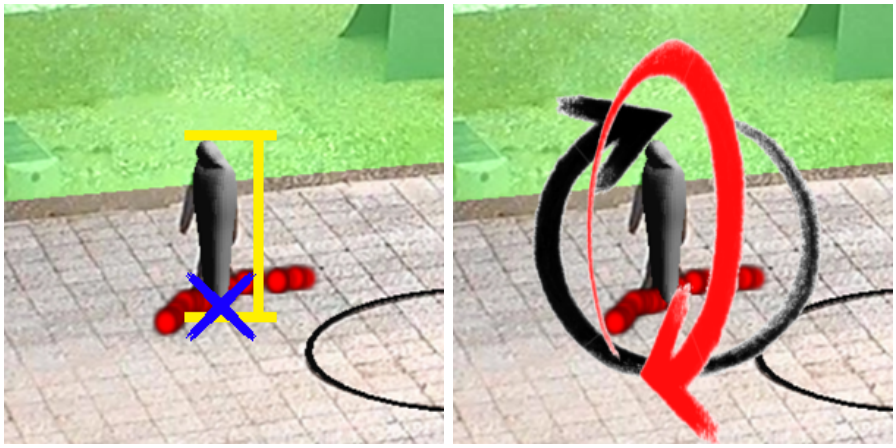


Figure 4.1: 3D model parameters obtained from camera calibration a) position and height, b) pitch and roll angles

response to color not only because of the cameras hardware but also because of the scene luminance is one of, if not the main reason of errors in re-identification systems.

The second constraint is commonly required by most state-of-the-art re-identification algorithms [5, 19, 34, 80], since working on the silhouette and not on the image window (as more generic content-based retrieval algorithms do) reduces the presence of outliers in the extracted features.

The third is peculiar to the method proposed in this chapter, since working in the 3D space requires at least the possibility to have a more or less precise estimation of the orientation, position and scale factors of a standing pedestrian. If not available, camera calibration can be estimated by statistical normalization [138] or vanishing points estimation [126].

The main strength of the proposed model is to have:

- A punctual description of the appearance, that allows to exploit details of the person appearance in the re-identification process and to be robust against miss-segmentation,
- A description independent from the point of view very robust to changes of the cameras FoVs,

The overall framework consists of three different layers: (1) appearance and geo-

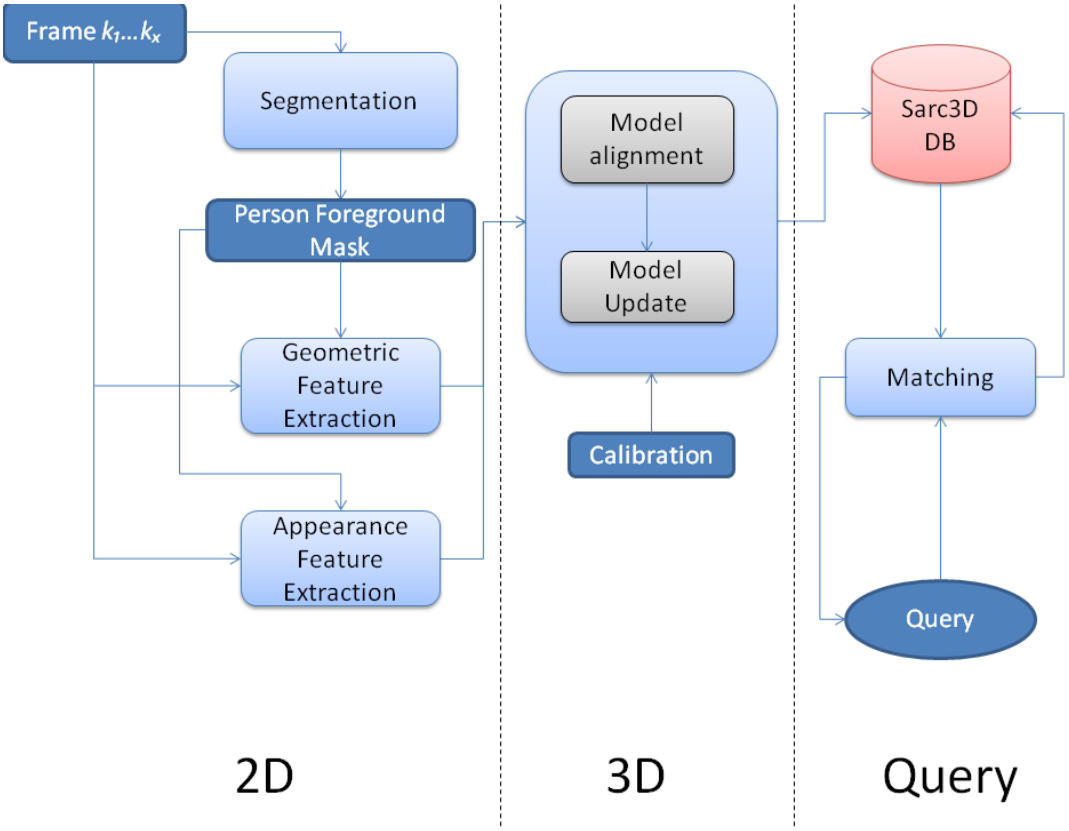
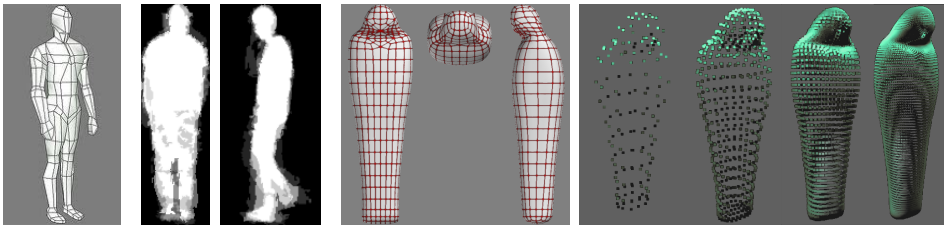


Figure 4.2: Schema of the proposed system

metric features extraction from the 2D images, (2) feature mapping on a 3D body model, and (3) model storage and query manager. A sketch of the system layers and internal modules is reported in Fig. 4.2. At the first stage, the set of video frames or still images containing different views of a person are provided to a segmentation module. If the samples are extracted from a video sequence captured with a still camera, a pixel-wise foreground mask of the person is estimated, otherwise a people detector [98, 196] is adopted possibly followed by a segmentation algorithms [31, 211] to select the person appearance Region of Interest (RoI). On the selected RoIs, a set of geometrical (e.g., feet position, person orientation, symmetries) and appearance (e.g., color histograms or texture) descriptors is then computed. These tasks are basically included on all the re-identification methods proposed in the past, with the aim of generating a signature for each person. Our specific choices will be discussed in the following sections.



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

Figure 4.3: (a) a human 3D model, (b) average silhouettes used for the model creation, (c) our simplified human model, (d) Different sampling densities of the SARC3D model used in our tests

The second layer is devoted to the construction and update of 3D model, which constitutes the main novelty of the proposed framework. The geometrical information are used to align the current view with respect to a 3D body model on which the appearance descriptors are mapped (see Sec. 4.2). The model allows a view independent representation of a single shot as well as the correct fusion of multiple views.

In the last layer, each computed model is stored and/or matched against all the reference items previously computed.

4.2 3D model based re-identification: the SARC3D Model

SARC3D is a vertex-based body model specifically developed for person re-identification. It is inspired by mesh models commonly used in computer graphics. The model has a fixed geometry and the relative vertex positions have been previously defined from a generic human shape manually constructed from real data. To this aim, side, frontal and top views of generic people were extracted from various surveillance videos; thus, an average silhouette has been computed for each view and manually transferred on a graphical 3D body model (see fig. 4.3(b)) exploiting standard 3D modeling tools and computer graphics techniques.The final product is a sarcophagus-like (fig. 4.3(c)) body hull.

A vertex set $\mathbf{v} = \{v_i\}, i = 1 \dots N_v$ is regularly sampled from the sarcophagus

surface, where the number N_v of vertices should be selected accordingly to the required resolution. This vertex set represents the final SARC3D model. The sampling is performed using a regular recursive sampling algorithm: the bounding box of the 3D object is recursively partitioned in an octree style, the center of each subdivided bounding box is considered, if it's distance to the surface is smaller than a given threshold, it's projection on the model surface is retained. The regular sampling guarantees that each triangle of the mesh has similar area, thus the same geometric weight for each vertex can be used when computing distances among models.

In several tests on real surveillance setups the surface was sampled obtaining a set of N_v vertices; small variations in the number of samples do not vary the performances of the model. Larger variations, instead, led to worse performance due to over or under fitting issues, as shown by preliminary tests. The tested samplings are shown in Fig.4.3(d), the four sampling corresponds respectively (from left to right) to sets of $N_v = 153$, $N_v = 628$, $N_v = 2026$ and $N_v = 10018$; Tests on the four samplings were performed on the ViSOR dataset. The lowest and highest densities ($N_v = 153$ and $N_v = 10018$) achieved very low accuracies (58% and 61% respectively), sampling the model with $N_v = 2026$ vertices achieved better performance (81%), while sampling $N_v = 628$ vertices produces the best results in terms of accuracy, resulting in 92% correct matches. Additionally, $N_v = 628$ is a good trade-off between speed and efficacy.

For each vertex v_i , the direction of the vector $\vec{n}_i$ normal to the 3D surface has been computed and it's stored together with the model[1].

The appearance part of the model is specific for each person. Instead of providing a texture definition for each triangle as in computer graphics approaches, a visual descriptor $\Xi_i{}^p$ is assigned to each vertex v_i of the p-th model and it is composed by color and/or texture features, as described in the section 4.2.1. Each person p is then characterized by a representative signature Γ^p defined as:

$$\Gamma^p = \{[\mathbf{v_i}], [\tilde{\mathbf{n}}_\mathbf{i}], [\mathbf{\Xi_i^p}], \mathbf{h^p}, \mathbf{x^p}, \theta^\mathbf{p}\} \tag{4.1}$$

Where v_i and $\vec{n}_i$ (respectively the vertices set and corresponding vertex nor-

[1]The SARC3D model is available at http://imagelab.ing.unimore.it/3DPeS/SARC3D.dae

mal) are the same for every instance of a SARC3D model, up to geometric transformation of rotation, translation and scale. $\Xi_i{}^p$ the visual descriptor, one for each vertex v_i and specific to each instance p of the model. h^p is the height of a person, used as scale factor for the SACR3D model, while $\mathbf{x^p}$ and θ^p are the position and orientation of the SARC3D model instance relative to the person p.

4.2.1 Color and Texture Feature Set

The visual descriptor Ξ_i[1]. of each vertex v_i is computed and updated by projecting the 3D vertex onto the 2D appearance image, after an adequate 2D to 3D alignment. Appearance information are then extracted from the image and stored.

The content of the vertex descriptor Ξ_i directly influences the re-identification performances. Two opposite requirements guide the selection: from one side, the vertex descriptor should be discriminative enough to avoid false matches; from the other side, over-fitting issues could arise due to noise and errors in the model alignment. For this reason, firstly as many descriptors as the model vertices are provided in order to obtain a punctual and localized description of the person 3D appearance, accounting to the discriminant details of each individual, on the opposite side the visual information of the vertexes neighborhood region is integrated at vertex level in order to avoid over-fitting.

Different types of descriptors have been tested, from mean colors to covariance matrices, but only histograms have produced meaningful results, which are reported in Section 6.2.1. Among the others, the following features and their combination have been tested as vertex descriptors for SARC3D:

- HSV color histogram; normalized histograms with 8 bins for the H channel, 4 bins for the S and V ones;
- RGB color histogram; normalized histograms with 8 bins for each channel have been adopted;
- HoG (Histograms of Gradients); normalized histograms with 9 bins;

[1]Superscript p will be omitted from now on for the sake of clarity

Figure 4.4: Extraction of color and texture features, a) original frame, b) vertices projection, c) highlighted region R_i of a random vertex, d) features extracted from region R_i

4.3 From 2D images to 3D models

To initialize and update the SARC3D model a suitable mapping from 2D image to the 3D model and vice versa is required. Lets assume that the pedestrian 2D shape has been segmented and the precise ground position and orientation is computed or estimated (see section 4.5), it is then possible to overlap the SARC3D model to the 2D appearance image of a walking person(see Fig. 4.8). The descriptor Ξ_i of each visible model vertex v_i is the initialized or updated with a set of appearance features (in our case, histograms $\mathbf{H_i}$) computed on a corresponding region $R_i$as depicted in Figure 4.4(c), following these steps for each vertex v_i of the model (also depicted in Fig. 4.4):

- The vertex v_i is projected on the image plane by exploiting the camera calibration, the estimated pedestrian orientation and perspective projection, obtaining the projected pixel position v'_i.
- A region of interest R_i is defined as the image patch centered on the vertex projected position v'_i of fixed size s_R ($s_R = 10$ in all our tests).
- Features are computed on the selected region R_i and assigned to the vertex descriptor Ξ_i.

The uniform sampling used to generate the vertex set guarantees that each descriptor corresponds to body surface areas of roughly the same expanse and

thus can be treated equally (i.e. with the same weight). However, since the 3D model is created from a single 2D image, some of the visual descriptors have seen the body surface from a frontal point of view, others from a lateral position and some of them are completely hidden (see Fig. 4.5).

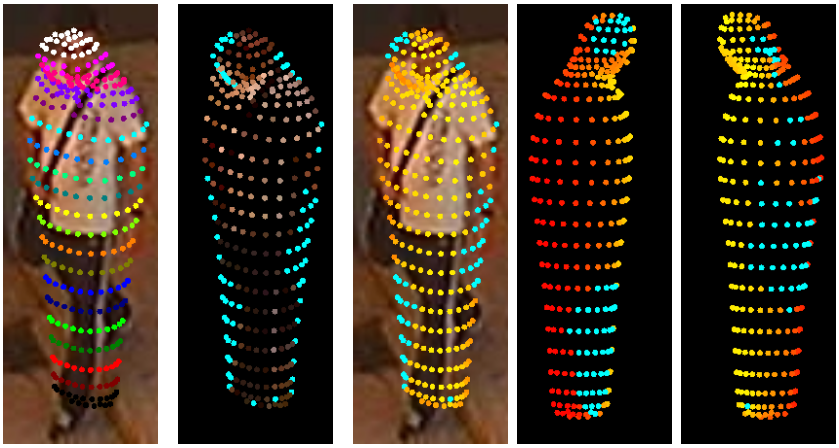
For this reason a reliability value ρ_i is computed and stored for each computed appearance feature. The reliability value takes into account how well and precisely the vertex descriptor has been captured from the data and it is computed as $\rho_i = \vec{n}_i \cdot \vec{p_I}$, where $\vec{p_I}$ is the normal vector of the camera image plane.

The reason behind the adoption of the dot product is that data from front-viewed points of the body surface and their surrounding surface are more reliable than that from lateral viewed ones. This reduces the drawbacks, due to errors in the model positioning and orientation, since stronger weights are assigned to the centrally-viewed points of the people appearance and lower weights to laterally-viewed ones, which are the most hit by misalignment.

The vertices belonging to the occluded side of the person are also projected onto the image and are initialized in the same way as the ones on the frontally-viewed side of the model, but their reliability has a negative value, due to the opposite directions of $\vec{n}_i$ and $\vec{p_I}$. In such a manner each vertex of the model could be initialized even with a single image: from a real view if available or using a sort of symmetry-based hypothesis in absence of information.

The vertices having no match with the current image (i.e. the vertices projected outside of the person silhouette) are also initialized: The vertices set is divided into 20 bands along the model height (see fig. 4.5(a), where the different bands are colored differently). After each model initialization or update, an average feature vector is computed for each band combining only the vertices correctly initialized via direct measure or symmetry. The vertices projected outside of the person silhouette (i.e. the vertices colored in light blue in the example in fig. 4.5) are then initialized with a copy of the average feature vector of their corresponding band and their reliability values are set to the minimum value (i.e., $\rho_i = 0$).

By means of the reliability value, vertices directly seen at least once ($\rho_i > 0$), vertices initialized using a mirroring hypothesis ($-1 \leq \rho_i < 0$) and vertices initialized from its neighborhood ($\rho_i = 0$) are distinguishable.



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

Figure 4.5: (a) SARC3D projection, (b) Color feature extraction, (c) Vertex reliability, (d,e) Left right views of the vertex reliability

The described steps of the initialization phase are depicted in Fig. 4.4.

4.4 Multi-view integration

As previously shown the model can be initialized and stored starting from a single person image, but as more images and views of the same person are available it can become more accurate and complete. As described in chapter 2 many recent re-identification methods exploit multiple views, often integrating them in a single multidimensional descriptor or averaging their values. Here thanks to the 3D model all the visual information available from multiple views can be exploited in a single reliable model at vertex level.

If multiple cameras are available or if the short-term tracking system provides more detections of the same person, the 3D model can be further refined and updated by integrating all the available frames. Two different integration methods have been adopted and tested: *histogram averaging* and *bag-of-histograms*.

Let $v_i{}^p$ and $v_i{}^s$ be homologous vertices of two SARC3D models (p and s) constructed from different views that need to be integrated into a new model q. Each vertex contains the feature sets $\Xi_i{}^p$ and $\Xi_i{}^s$ respectively, that will be fused into a new feature set $\hat{\Xi}_i$.

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

Figure 4.6: Various models created with the corresponding source images

In the first approach, each visual descriptor Ξ_i contains a single feature vector for each type of feature exploited (i.e. $\Xi_i = \{\mathbf{H_i}, \rho_\mathbf{i}\}$), each merged feature set is computed as the weighted average of the source ones as in Equation 4.2:

$$\hat{\mathbf{H}}_\mathbf{i} = \frac{{\rho_i}^p \mathbf{H_i}^p + {\rho_i}^s \mathbf{H_i}^s}{{\rho_i}^p + {\rho_i}^s}, \qquad \hat{\rho_i}^p = \frac{{\rho_i}^p + {\rho_i}^s}{2} \tag{4.2}$$

Where $\mathbf{H_i}^p$ and $\mathbf{H_i}^s$ are HSV, RGB or HoG histograms of the visual descriptors ${\Xi_i}^p$ and ${\Xi_i}^s$ respectively, and the absolute values of the vertex reliabilities (${\rho_i}^p$ and ${\rho_i}^s$) are used as weights.

In the second case, instead, each visual descriptor contains multiple instances of each feature, $\Xi_i = [\mathbf{H_{i,j}}, \rho_\mathbf{i,j}, \mathbf{t_{i,j}}]$ with $j = 1...n_{\Xi_i}$, where $t_{i,j}$ is a time-stamp assigned to the visual feature $\mathbf{H_{i,j}}$ and n_{Ξ_i} is the number of fused views, the new visual descriptor $\hat{\Xi}_i$ is then simply the union of the feature sets of ${\Xi_i}^p$ and ${\Xi_i}^s$.

Fig. 4.6 shows some example SARC3D models generated integrating multiple views by averaging the feature vectors.

4.4.1 View selection

Not all views should be used for the initialization and update of the model. Errors in the tracking step, noise and bad calibration could lead to degradation of the model. Additionally in the case when each visual descriptor contains multiple instances of each feature a "forget" mechanism need to be defined, in order to prevent the visual descriptors dimension to grow excessively. To this aim a rule based approach was implemented, composed by the following five rules,

(a) (b) (c)

Figure 4.7: Occlusion detection: (a) the input frame, (b) the aligned 3D models and (c) the mask generated by the rendering system. Since the blue and green objects are connected, the corresponding models are frozen and not updated during the occlusion

Occlusion check: In addition to 2D occlusion detection algorithms [204], a computer graphic based generative approach is used: for the selected camera view and for each person p visible from that camera, a binary image mask $\hat{I}_p$ is rendered using standard computer graphics techniques. Each time two model masks are overlapping or connected an occlusion is detected. To avoid false pixel to model assignments, both the occluding and the occluded models are not updated. A visual example of the 3D occlusion detection is shown in Fig. 4.7.

Model to foreground overlapping: The reliability of the model positioning could be evaluated considering the overlapping area between the 2D foreground mask $\hat{F}_p$ and the rendered images $\hat{I}_p$. For each person, the overlapping score o_p is computed as the ratio between the number of foreground pixels that overlap with $\hat{I}_p$ with respect to the total number of silhouette pixels. If o_p is higher than a strong threshold (e.g., 95% in our experiments) the selected view is marked as good. Otherwise the alignment is not precise enough or the person is not assuming a standing position compliant with the sarcophagus model.

Orientation reliability: The reliability of the orientation estimation is evaluated twofold: First the score of the orientation estimation (see Sec. 4.5.2) is considered. If higher than a predefined threshold the orientation estimated is considered reliable. Secondly, the sequence of the estimated orientations is evaluated: if the distribution of the differences between consecutive orientations has a high variance, the trajectory is not stable and the orientation becomes unreliable. If both conditions hold true, the estimated orientation is considered reliable.

Information gain: In order for the new view to be used, it should add meaningful informative content to the model. This is evaluated by considering the number of updated vertices v_i. A vertex v_i is considered updated if a new feature vector $\{\hat{\mathbf{H}}_\mathbf{i}, \hat{\rho}_i\}$ is added to it. The feature vector is added to the visual descriptor Ξ_i of v_i if either $\hat{\rho}_i$ of the new vector is greater than the optical reliability of each other feature vector in Ξ_i or the distance between the new feature vector and each other feature vector in Ξ_i is greater than a given threshold (see Sec. 4.6 for the definition of distance between feature vectors).
If all these conditions hold true, the estimated orientation and position is considered reliable and the selected view can be exploited to initialize (or update) the model.

Forget rule: A fifth rules has been defined that implements a "forget" mechanism in order to prevent the visual descriptors dimension to grow excessively. A feature vector $[\mathbf{H}_\mathbf{i}, \rho_\mathbf{i}, \mathbf{t}_\mathbf{i}]$ of Ξ_i is forgotten if the time distance between its timestamp t_i and the current date is great enough and if it's deletion do not reduce the informative content (as measured in the previous rule) of the model too much.

4.5 Model Alignment

The alignment of the 3D model on a camera image is the most critical step of the approach. Let us assume that the set of cameras are fixed and calibrated, as in many surveillance systems. Under this hypothesis, the model alignment can be split into two independent steps, i.e., *model placement* and *orientation estimation* (see Fig. 4.8).

4.5.1 Model Placement

The model placement is obtained by estimating the position of the person feet on the ground plane and it is provided by means of a homography transformation [44] of a selected image point. The person is also assumed to be in a vertical standing posture, which also defines two of the three angles required for the model alignment thanks to the camera calibration. Usually a background segmentation step is exploited in order to detect and extract the bounding box of a person

Figure 4.8: Positioning and orientation of the SARC3D model

silhouette and the silhouette itself from a single camera view. The feet position is estimated to be the midpoint of the bottom part of the bounding box or the position of the lowest point of the silhouette. If tracking is available, quadratic function fitting can be used in order to filter out and smooth the person feet position estimation. Tracking also allows providing multiple consecutive images for the model creation.

A detection and tracking method was developed [203] and exploited in combination with SARC3D. This method is composed by two layers: first background subtraction is performed in order to detect moving pedestrians and extract their foreground appearance image, and then particle filters are exploited in order to track their movements inside a single camera view.

Appendix A report another detection and tracking method successfully integrated with SARC3D as a result of a collaboration with Ákos Utasi, Csaba Benedek, Tamás Szirányi of the Computer and Automation Research Institute, Hungarian Academy of Sciences and presented in [22].

4.5.1.1 Background Suppression

Since in the regarded application fixed cameras are usually exploited, every frame is processed with a background subtraction system. The background subtraction algorithm employed works similarly to the one presented in [59]: A difference image DB_t is computed between the current frame I_t and a background model

B_t. A threshold with hysteresis is adopted, points in DB_t are selected if greater than a low threshold, then morphological operator (closing and opening) are applied on the image in order to eliminate isolated points or very small spots due to noise. Then labeling is performed, by accepting blobs containing at least one point greater than a high threshold. The labeled image F_t at time t thus contains a number of Foreground Blobs. On the segmented foreground blobs additional analyses are performed, consisting of two steps: 1) blobs with an area less than a threshold (which depends on the distance between camera and scene and on the typical size of objects) are discarded; 2) the average optical flow is computed for each blob. If it's smaller than a given threshold they are discarded. They are considered as apparently moving objects (due to a locally wrong background) and thus not accepted as real detection of people.

The model of background B_t is based on the statistical assumption that the background points should be the most probable points observed in a finite window of observation time. As statistical function the median of sampled frames is exploited, an adaptive function which takes into account past background values and the current frame is also employed. Points belonging to detected person are not taken into consideration during the background model estimation. At initialization time, the background model is initialized with a novel algorithm based on the recursive Hadamard transform. The new algorithm is detailed in Appendix B.

4.5.1.2 People Tracking

Tracking is performed on the foreground image F_t with the use of particle filters.

Goal of a single camera tracking system is to estimate the state $\mathbf{x}_t$ at frame t given the set of observations $\{\mathbf{z}_{1\ldots t}\}$. At this stage it is assumed for simplicity that each tracked object is moving regardless of other people in the scene. Thus, an independent tracker is applied for each object in order to estimate its state, which is assumed to be composed by the coordinates of the gravity center.

$$\mathbf{x} \doteq \{x_c, y_c\}\,. \tag{4.3}$$

In a probabilistic framework, $\mathbf{x}$ is a random variable, with the associated prob-

ability density function $p(\mathbf{x}|\mathbf{z})$ which is the objective of our estimation process. The most likely position $\hat{\mathbf{x}}$ of each tracked person can be estimated at time t from the relative pdf as $\hat{\mathbf{x}}_t = E(\mathbf{x}_t)$.

To this aim a generic particle filtering technique is adopted; for the sake of completeness the base equations using the notation proposed by Arulampalam et al. [13] in their famous tutorial are reported here.

Let $\{\mathbf{x}^i, w^i\}_{i:1..N}$ be a characterization of the posterior probability $p(\mathbf{x}|\mathbf{z})$, where $\{\mathbf{x}^i, i = 1..N\}$ is the set of support points (particles) with their weights $\{w^i, i = 1..N\}$. The weights are normalized to add up to one. By means of this set of weighted particles the posterior probability can be approximated as:

$$p(\mathbf{x}\,|\mathbf{z}) \approx \sum_{i=1}^{N} w^i \delta\left(\mathbf{x} - \mathbf{x}^i\right) \tag{4.4}$$

As importance density the prior $p(\mathbf{z}_t|\mathbf{x}_t^i)$ is adopted, so that the particle weights can be iteratively estimated as:

$$w_t^i \propto w_{t-1}^i \cdot p\left(\mathbf{z}_t|\mathbf{x}_t^i\right) \tag{4.5}$$

and the resampling step is executed only if the measure of degeneracy $\widehat{N_{eff}}$ (Eq. 4.6) is lower than a threshold N_e.

$$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N} \left(w^i\right)^2} \tag{4.6}$$

The particle filter tracking can be summarized as a 5 steps process: after initialization of the particle system, the new particle filter state is predicted (i.e. the particles positions), observations are taken (measurements, likelihoods), through the observations the new state estimate is updated and finally particles are resampled. The method is summarized by the two pseudo-code algorithms reported in Fig. 4.9 (from Algorithms 2 and 3 in [76]). The adopted likelihood function and the new occlusion-based process model, instead, are fully described in the following.

RE-SAMPLING ALGORITHM

- Initialize the CDF: $c_1 = 0$
- For $i = 2 : N_s$
  - Construct CDF: $c_i = c_{i-1} + w_k^i$
- Start at the bottom of the CDF: $i = 1$
- Draw a starting point: $u_1 \sim \mathcal{U}[0, N_s^{-1}]$
- FOR $j = 1 : N_s$
  - Move along the CDF: $u_j = u_i + N_s^{-1}(j-1)$
  - WHILE $u_j > c_i$
    - i = i+1
  - Assign sample: $\mathbf{x}_k^{j*} = \mathbf{x}_k^i$
  - Assign weight: $w_k^j = N_s^{-1}$

GENERIC PARTICLE FILTER

- FOR $i = 1 : N_s$
  - Draw $\mathbf{x}_k^i \sim q(\mathbf{x}_k | \mathbf{x}_{k-1}^i, \mathbf{z}_k$
  - Assign the particle a weight, w_k^i according to Eq. (4.5)
- Normalize w_k^i such that $\sum_i w_k^i = 1$
- Calculate $\widehat{N_{eff}}$ using Eq. (4.6)
- IF $\widehat{N_{eff}} < N_T$
  - Re-sample using RE-SAMPLING ALGORITHM

Figure 4.9: Pseudo-code of re-sampling and particle filter algorithms

The likelihood function Similar to work of Li et al. [128], for each object a set of three appearance models is stored and kept updated $AM_t = \{am_t^1 \doteq Fx, am_t^2 \doteq A_t, am_t^3 \doteq D_t\}$. Pixels are assumed independent of each other. The fixed model Fx_t contains the appearance of the tracked object stored at the initialization phase. The Adaptive model A_t stores the mean of N object appearances sampled at regular time steps. Finally, the Dynamic model D_t is estimated averaging D_{t-1} with the current frame.

All these models can take advantage of a foreground segmentation of the current frame. Only the image points classified as foreground pixels concur to the model estimation. Calling I_t the current frame, the model update equations are:

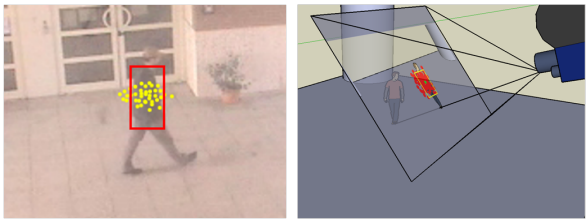
$$\begin{aligned} AM_t &= \{am^1, am^2, am^3\} \\ &\begin{cases} am^1 = Fx(\mathbf{x}) = I_{t_1}(\mathbf{x} - \hat{\mathbf{x}}_\mathbf{t}) \\ am^2 = A_t(\mathbf{x}) = \frac{1}{N} \cdot \sum_{j=1}^{N} I_{t-j\Delta t}(\mathbf{x} - \hat{\mathbf{x}}_{\mathbf{t-j\Delta t}}) \\ am^3 = D_t(\mathbf{x}) = \frac{1}{2} \cdot D_{t-1}(\mathbf{x}) + I_t(\mathbf{x} - \hat{\mathbf{x}}_t) \end{cases} \end{aligned} \tag{4.7}$$

where $I_{t_1^i}$ is the first frame of the i-th object, i.e. when it is entered the scene; the notation $\mathbf{x} - \hat{\mathbf{x}}_\mathbf{t}$ is used to indicate the frame by frame alignment of the model using the estimated position of the target $\hat{x}_t$.

On the whole, the state $\{\hat{x}, AM\}$ of a tracked object is composed by $\hat{x}$, i.e., the position of the center of mass in the image plane (estimated by particle filtering), and its appearance AM (updated using Eq. 4.7).

With the likelihood function$p(\mathbf{z_t}|\mathbf{x_t})$ can be stimated, i.e., how likely a particular object position is to produce the current frame. Practically, the estimated models for each object at the previous frame AM_{t-1} must be compared with the current image I_t. A pixel-wise comparison usually leads to errors if the model is not exactly aligned with the current frame, since a pixel-by-pixel distance is not a monotonic function. Therefore the distance measure between the model and the current observation is based on aggregate functions like color histograms.

Let be $R_d\,(I, \mathbf{x})$ a rectangular region of the image I, centered on the point $\mathbf{x}$, fixed shape ratio and scale factor s_R. Let $H(\cdot)$ be the histogram of the image argument. Then, at sampling time one of the three appearance models $am^j \in AM_t$



(a) (b)

Figure 4.10: Graphical representation of the tracked state through particle filter. Dots are the particles positions and the rectangle is the ROI for the likelihood computation and model update

is randomly selected for each particle. The likelihood value $p(\mathbf{z}_t|\mathbf{x}_t)$ is extracted from a zero-mean normal distribution using the Bhattacharyya distance $\Phi(\cdot)$ between the color histogram from the current image $H\left(R_d\left(I_t, \hat{x}_t\right)\right)$ and the selected model histogram $H(am^j)$ as in Eq. 4.8

$$\begin{aligned} p(\mathbf{z}_t|\mathbf{x}_t^i) &= p(\mathbf{x}_t|\mathbf{x}_t^i, am^j) = \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{\Phi\left(H(R_d(I_t,\hat{x}_t)),H(am^j)\right)}{2\sigma^2}} \end{aligned} \tag{4.8}$$

The Process Model The motion of a person in a scene is difficult to predict and is seldom linear, in particular if the object position is measured in image coordinates. Therefore, usually the random walk equations are adopted for the state prediction step [82]. Our approach is slightly different and takes into account the previous movement of the person in addition to the Gaussian noise:

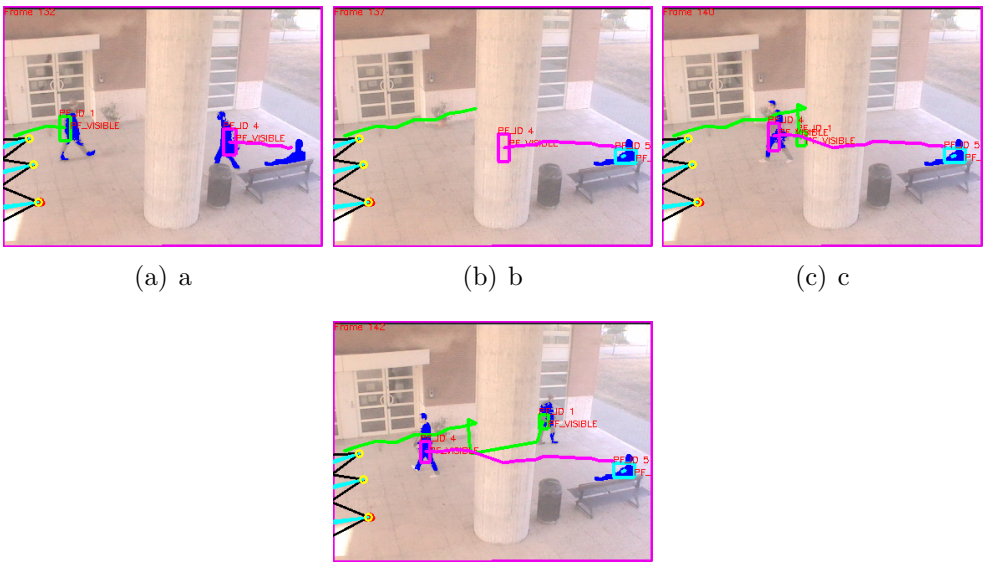
$$\tilde{\mathbf{x}}_t^i = \mathbf{x}_{t-1}^i + \frac{\vec{v}_{t-1} + 2\vec{v}_{t-2}}{3} + \mathcal{N}(0, \sigma I_2) \tag{4.9}$$

where $\vec{v}_t = \hat{x}_t - \hat{x}_{t-1}$. The Gaussian noise of Eq. 4.9 should take into account the nonlinear nature of the human motion. Furthermore the tracker can be misled by occlusions and shape changes. Thus, the Gaussian noise should be large enough to manage the unpredictable changes in speed and direction, but it does not set to naught the linear prediction. To this aim a spherical covariance is exploited, with a dynamic parameter σ which depends on the likelihood score computed at

the estimated position $\hat{\mathbf{x}}$:

$$\sigma_t = \frac{\sigma_0}{(1 + \alpha p(\mathbf{z}_t|\hat{\mathbf{x}}_t))}. \tag{4.10}$$

where σ_0 is mandatory since the variance should be greater than zero and α is a predefined constant. During occlusions or quick motion changes the distance function computed in the estimated position grows; increasing the noise term. This will assure that particles will be much spread in the next step. Figure 4.11 reports an example of long-lasting occlusions. Even that, the tracker system can manage the label assignments after the occlusions thanks to the dynamic nature of the noise term.



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

(d) d

Figure 4.11: Example of person tracking during a strong occlusion

Initialization As above mentioned, each object is independently considered and tracked; every time a person enters the scene a new particle filter should be created and initialized. To this aim the foreground mask extracted from the background subtraction algorithm is labeled and unassigned blobs are classified as new objects. A blob is unassigned if no particle has required that blob for the likelihood function estimation. To avoid wrong assignments between new blobs and particles associated with other objects, a blob is associated to a particle (and

then eliminated from the list of new objects to be tracked) if the correspondent Bhattacharyya distance is under a threshold.

The initial position $\hat{x}_0$ of the object is set equal to the blob gravity center and the three appearance models are initialized with the blob appearance.

4.5.2 Orientation Estimation

The last parameter to estimate is the orientation of the model with respect to its vertical axis. To this aim, two different approaches are available in the system. If the detection comes from a video sequence and the person trajectory has been provided through a tracking algorithm, the orientation could be inferred from the trajectory itself. By assuming that people move forward the trajectory on the ground plane could be exploited to give a first approximation of the orientation. Given a detected person, a window of K frames is considered and the corresponding trajectory on the ground plane. A quadratic curve is then fitted on the trajectory and the fit score is used as orientation reliability. If it is above a predefined threshold, the final orientation is generated from the curve tangent. In fig.4.12(a) and 4.12(b) a sample frame of the corresponding model placement and orientation is provided. In particular, the sample positions used for the curve fitting and orientation estimation are highlighted.

(a) (b)

Figure 4.12: (a) A frame from a video, (b) Automatic 3D positioning and orientation

The previous method works only when tracking is available; in order to estimate the correct orientation of a person from a single image a new system was developed [25] based on machine learning techniques and mixtures of approximated wrapped gaussians. It's a new and very general approach which exploits state of

the art descriptors and detectors, but ensembles them in a unique angle-oriented classifier. Since the body orientation is mainly related to shape and edges, the best features are straightforwardly related to luminance gradients, without the influence of colors. Histograms of Oriented Gradients (HoG) features [62] are adopted. Orientations are quantized into 8 different classes (See Fig. 4.14). The main orientations are singularly recognized with an array of Extremely Randomized Trees classifiers [87], which proved to be very fast and powerful in this case. Moreover, the detectors response are integrated in a single probability density function generated as a Mixture of Wrapped Distributions, and in particular as a Mixture of Approximated Wrapped Gaussian (MoAWG) weighted by the detector outputs. The maximum of this probability density function is the answer of the orientation problem that is further quantized in the main directions, for filtering errors and noise and for making an easy comparison with ground truth data. Fig. 4.13 outlines the proposed orientation detection system.

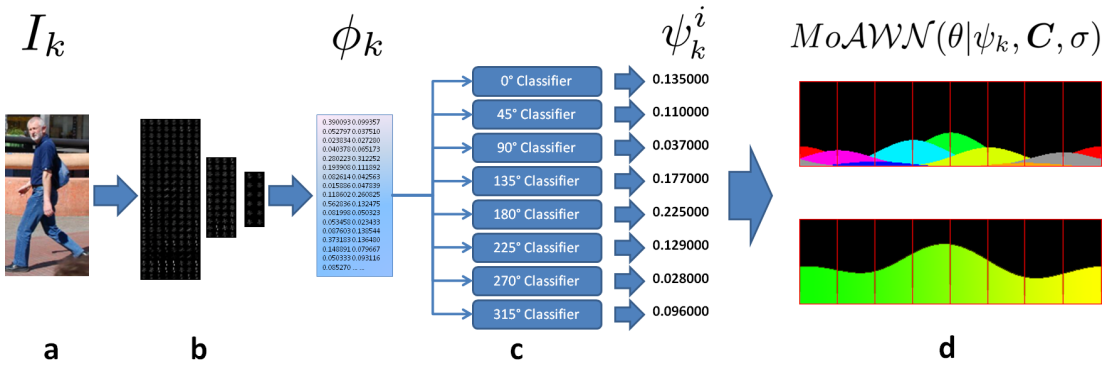


Figure 4.13: A schema of the proposed method, (a) Input image, (b) Multi-Level HoG, (c) Array of classifiers, (d) the Mixture of Approximated Wrapped Gaussians

As a matter of fact, the estimation of people orientation is an intrinsically continuous problem. The discretized classes are not well separated and sometimes even overlapped (due to the torsion movements of the body). The MoAWG acts as "interpolation" of the outputs of different trained classifiers: each binary classifier is trained to select a positive region of the feature space, which is not related and not imperatively disjoint with the others. The final step (AWG) integrates different contributions and thus improves the classification when the

orientation is quite ambiguous.

The resulting approach is fast, simple and very effective, it can be generalized for a different number of main directions, and it can be adopted in many other problems where main directions must be recognized. In Section 6.4 several tests and comparisons of the orientations detection system only are performed, on different publicly available datasets, such as the TUD Multiview Pedestrians dataset [11], SARC3D[23] and 3DPES[24], both available on OpenVisor[1], and video sequences from PETS. Comparisons with previous methods demonstrated very satisfactory improvements of about 18% with respect to the state of the art, achieving up to 65% of accuracy on the TUD Multiview Pedestrians dataset using eight directions and more than 80% of accuracy using four directions.

4.5.2.1 System overview

The proposed system aims at estimating the orientation of a person with respect to the camera point of view. The input is a single detection I_k of a person, obtained cropping an image or a video frame on the bounding box provided by a generic appearance based people detector [62, 73]. Since motion or background information is neglected, the person silhouette is not available. The orientation $\theta_k \in (-\pi, \pi]$ of the person is the angle between the main direction of the person and the horizontal axis of the image plane. The precise value of θ_k is ambiguous and impossible to measure since head, shoulders and legs could be differently oriented. Thus, a discrete set of directions instead of a continuous range of values is more appropriated. Let us define the set $\boldsymbol{C}$ of N discrete orientations sampled from the interval $(-\pi, \pi]$ as

$$\begin{aligned} \boldsymbol{C} &= \{c^i\}, i \in (0, N-1), \\ c^i &= \left(\left(\tfrac{2\pi i}{N} + \pi\right) \mod 2\pi\right) - \pi. \end{aligned} \tag{4.11}$$

In our experiments N was set to $N = 8$, obtaining the eight main directions depicted in Fig. 4.14. For each class c^i, a specific binary classifier was trained on a set of HoG descriptors [62]; a classification score ψ^i instead of a boolean response is also required. A first orientation estimation $\bar{c}_k$ of the image I_k could

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

be directly obtained from the outputs $\Psi_k = \{\psi_k^1, \ldots, \psi_k^N\}$ of the classifiers:

$$\bar{c_k} = c^j, j = \arg\max_i \psi_k^i. \tag{4.12}$$

(a) (b)

Figure 4.14: The eight directions recognized by the proposed system and the corresponding color labels

Using Eq.4.12 the estimated orientation does not take into account the output of all the classifiers, but the winner one only. In particular, due to the ambiguity of the human direction above described, more than one classifier could positively react and a more precise estimation of the main orientation could be obtained by combining the results of all the classifiers. To this aim, the continuous distribution $p(\theta|I_k)$ is estimated as a function of the classifier outputs. The person orientation $\bar{\theta_k}$ and its corresponding discretized class $c(\bar{\theta_k})$ are now computed maximizing the previous distribution:

$$\bar{\theta_k} = \arg\max_{\theta \in (-\pi, \pi]} p(\theta|I_k). \tag{4.13}$$

Algorithms 1 and 2 report a pseudo-code description of the classification steps, while the following subsections will detail the set of classifiers and the integration of their outputs using a circular statistic approach.

Algorithm 1 Discrete Orientation Classifiers

Require: $N, \Psi = \{\Psi^1, \ldots, \Psi^N\}$, set of trained classifiers

function MULTILEVELHOG(I)
  $\{Q_j^1\}, j = 1 \ldots 192 \leftarrow$ SPLIT($I, 8, 24$) ▷ Level 1
  $\phi_j^1 = HoG(Q_j^1)$
  $I \Leftarrow$ RESIZE($I, 0.5$)
  $\{Q_j^2\}, j = 1 \ldots 48 \leftarrow$ SPLIT($I, 4, 12$) ▷ Level 2
  $\phi_j^2 = HoG(Q_j^2)$
  $I \Leftarrow$ RESIZE($I, 0.5$)
  $\{Q_j^3\}, j = 1 \ldots 12 \leftarrow$ SPLIT($I, 2, 6$) ▷ Level 3
  $\phi_j^3 = HoG(Q_j^3)$
  $\phi = [\phi_1^1 \ldots \phi_{128}^1 | \phi_1^2 \ldots \phi_{48}^2 | \phi_1^3 \ldots \phi_{12}^3]$
  Normalize(ϕ)
  return ϕ
end function

function FINDORIENTATIONS(I)
  $\{I_k\} \leftarrow$ PEOPLEDETECTOR(I)
  for $k = 1 \rightarrow K$ **do**
    $\phi_k \leftarrow$ MULTILEVELHOG(I_k)
    for $i = 1 \rightarrow N$ **do**
      $\psi^i \leftarrow \Psi(\phi_k)$
    end for
    $\bar{c_k} \leftarrow \arg\max_i \psi_k^i$.
  end for
end function

4.5.2.2 Discrete Orientation Classifiers

For each detected person, a 2268-dimensional feature vector is computed based on the HoG descriptor[62]. The color image cropped around the person is firstly converted into a single channel image; the first direction of a Principal Component Analysis space reduction is selected. With respect to the luminance channel, the PCA-based image channel preserves and even enhances the edge gradients. Thus, a multi-level HoG feature vector is computed, dividing the input image into blocks at three different levels: the first level contains 8x24 non-overlapping blocks, the second level 4x12 blocks and the last level 2x6 blocks. At each level the image is down-sampled with a scale factor of 0.5. A histogram of oriented

gradients quantized in 9 wrapped bins is computed on each of the 252 blocks and normalized over 2x2 sets of blocks. During the histogram computation, the tri-linear interpolation described in [62] is preserved. The 2268-dimensional feature vector ϕ_k is obtained concatenating the 9 histogram values of the 252 blocks computed over I_k; ϕ_k acts as appearance descriptor of the images and it is sent to the array of classifiers.

Due to the very high dimensionality of the input feature vector, the Extremely Randomized Trees classifiers introduced by Geurts *et. al.*[87] are adopted. The Extremely Randomized Trees are similar to Random Trees but instead of using bagging selection they keep the same input training set to train all the trees. For binary classification problems only, Random Trees allow the estimation of a fuzzy-predicted class label, i.e., a confidence value of the binary classification result. In our case, given a feature vector ϕ_k, each of the N classifiers provides a value $\{\psi^i, i = 1, \ldots, N\}$ calculated as the proportion of decision trees that classified the input to the winner class. A discrete label of the image orientation could be generated using Eq. 4.12.

4.5.2.3 Output Filtering by Circular Statistics

Instead of directly using the outputs of the N trained classifiers to generate a discrete class label, the classification results are integrated in a continuous probabilistic distribution $p(\theta|I)$. The reason of this step is mainly due to the overlapping of the orientation classes, which leads to have more than one high response from the set of discrete-orientation classifiers.

The terms ψ^i are used as weights of a mixture of wrapped distributions, each centered on the N selected orientations θ_i. Directional statistics has been widely studied in the past and Wrapped Gaussians or the most general von Mises distributions are the widest adopted models [143] to manage periodic data such as angles [2]. The probability density function of a wrapped normal distribution is

$$\mathcal{WN}(\theta|\theta_0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \sum_{k=-\infty}^{+\infty} e^{-\frac{(\theta-\theta_0+2k\pi)^2}{2\sigma^2}}, \tag{4.14}$$

where μ and σ^2 are the corresponding means and variance. A very interesting

approximated version of the Wrapped Gaussian has been presented by Bahlmann in [17] to deal with semi-periodic multivariate data in handwritten character recognition and successively used in [45] for trajectory description and clustering. The corresponding probability density function is

$$\mathcal{AWN}(\theta|\theta_0,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{((\theta-\theta_0) \mod 2\pi)^2}{2\sigma^2}}. \tag{4.15}$$

A mixture of $\mathcal{AWN}$ is obtained as a weighted sum of $\mathcal{AWN}$ probability density functions:

$$MoA\mathcal{WN}(\theta|\boldsymbol{w},\boldsymbol{\theta}_0,\boldsymbol{\sigma}) = \sum_{i=1}^{N} w_i \cdot \mathcal{AWN}(\theta|\theta_{0,i},\sigma_i) \tag{4.16}$$

The required function for the orientation estimation is thus obtained using a Mixture of Approximated Wrapped Gaussian as in Eq. 4.16:

$$p(\theta|I_k) = MoA\mathcal{WN}(\theta|\psi_k,\boldsymbol{C},\sigma) \tag{4.17}$$

where the variance σ was set to a fixed value for all the components. The σ parameter of Eq. 4.17 depends on the number of adopted classifiers: if σ is 0 the AWG step is disabled. Increasing the σ value includes in the final response the contributions of more neighbor classifiers.

The person direction is estimated using Eq. 4.13 through Mean-Shift optimization with starting seeds on the c^i values. Fig. 4.13 shows all the steps of the proposed method and in particular the final filtering step obtained with the Mixture of Approximated Wrapped Gaussians which is also described in Algorithm 2.

4.6 Distance metric for People re-identification

Since the signature of the p-th person instance is composed by the set of vertex descriptors $\Gamma^p = \{{\Xi_1}^p \dots {\Xi_N}^p\}$, the distance $D_H(\Gamma^p,\Gamma^q)$ between two models Γ^p and Γ^q could be decomposed as the weighted average of the vertex-wise visual descriptors Ξ_i distances. The product of the vertex reliabilities could be used as weights in order to emphasize the visible parts of the models.

Algorithm 2 Orientation estimation with a Mixture of Approximated Wrapped Gaussians

Require: $N, \Gamma = \left\{\Gamma^1, \ldots, \Gamma^N\right\}$, set of trained classifiers

function FINDORIENTATIONS2(I)
  $\{I_k\} \leftarrow$ PEOPLEDETECTOR(I)
  for $k = 1 \rightarrow K$ **do**
    $\phi_k \leftarrow$ MULTILEVELHOG(I_k)
    for $i = 1 \rightarrow N$ **do**
      $\psi^i \leftarrow \Gamma^i(\phi_k)$
    end for
    $p(\theta|I_k) \leftarrow MoA\mathcal{WN}(\theta|\psi_k, \boldsymbol{C}, \sigma)$
    $\theta_k \leftarrow \arg\max_{\theta \in [-\pi \ldots \pi]} p(\theta|I_k)$ ▷ Mean Shift Maximization
    if Continuous Output **then**
      return θ_k
    else
      $i = \left(\theta_k + 2\pi \cdot \frac{N}{2\pi}\right) \mod N$
      return c_i
    end if
  end for
end function

$$D_H(\Gamma^p, \Gamma^t) = \frac{\sum_{i=1\ldots M} \left(w_i \cdot d(\Xi_i^p, \Xi_i^q)\right)}{\sum_{i=1\ldots M} \left(w_i\right)} \tag{4.18}$$

where

$$w_i = \left| abs(\rho_{i,j}^p) \cdot abs(\rho_{i,j}^q) \right| \tag{4.19}$$

The vertex distance $d(\Xi_i^p, \Xi_i^q)$ is strictly related to the adopted feature type. Usually the Hellinger distance is applied to compare color histograms. For example, the distance

$$d(\Xi_i^p, \Xi_i^q) = \min_j \left(d_{He}\left(\mathbf{H_{i,j}^p}, \mathbf{H_{i,j}^q}\right)\right) = \sqrt{1 - \sum_{r,g,b} \sqrt{H_{i,j}^p(r,g,b) \cdot H_{i,j}^q(r,g,b)}} \tag{4.20}$$

is the distance between two feature vectors belonging to the i-th vertex of two SARC3D models Γ^p and Γ^q. Each feature vector contains a bag of descriptors (in this case a bag of color histograms $\mathbf{H_{i,j}^p}$ with $j = 1 \ldots n$ where n is the number of

color histograms in a vertex bag of descriptors)

When color and gradients histograms are combined, the corresponding distance measure is:

$$d(\Xi_i^p, \Xi_i^q) = \min_j \left(d_{He}\left(\mathbf{H_{i,j}^p}, \mathbf{H_{i,j}^q}\right) + d_{Hg}\left(\mathbf{Hg_{i,j}^p}, \mathbf{Hg_{i,j}^q}\right)\right) \tag{4.21}$$

where d_{Hg} is the distance between gradients histograms, defined as:

$$d_{Hg}\left(\mathbf{Hg_{i,j}^p}, \mathbf{Hg_{i,j}^q}\right) = \sqrt{\sum_a \left(\mathbf{Hg_{i,j}^p}(\mathbf{a}) - \mathbf{Hg_{i,j}^q}(\mathbf{a})\right)^2} \tag{4.22}$$

4.6.1 Vertex saliency for detail-oriented re-identification

This generic global distance assumes that each vertex has the same importance and the weights w_i are based only on optical properties of the projections or the reliability of the data. Global features are useful to reduce the number of candidates or if the resolution is low. However, the final decision should be guided by original patterns and details, as humans normally do to recognize people without biometric information (e.g., a logo in a specific position of the shirt). To this aim the vertex feature vector Ξ_i is enriched with a saliency measure $\varsigma_i^p \in [0 \dots 1]$. Given a set of body models, the saliency of each vertex is related to its minimum distance from all the corresponding vertices belonging to the other models:

$$\varsigma_i^p \propto \min_t \left(d_H(\mathbf{H_i^p}, \mathbf{H_i^t})\right) + s_0, \qquad \sum \varsigma_i^p = 1 \tag{4.23}$$

where ς_0 is a fixed parameter that gives a minimum saliency to each vertex. If ς is low, the vertex appearance is not distinctive; otherwise, the vertex has completely original properties and it could be used as a specific identifier of the person. The corresponding saliency-based distance D_ς can be formulated based on new weights by substituting $\prime w_i$ to w_i in eq.4.19.

$$w_i' = abs(\rho_i^p) \cdot abs(\rho_i^q) \cdot \varsigma_i^p \cdot \varsigma_i^q \tag{4.24}$$

This saliency-based distance D_ς cannot replace Eq. 4.18, since it focuses on

details discarding global information and then leading to macroscopic errors; the re-identification should be based on both global (D_H) and local (D_ς) similarities. Thus, the final distance measure $D_{H\varsigma}$ used for re-identification is the product of the two contributions $D_{H\varsigma} = D_H \cdot D_\varsigma$.

Chapter 5

From SARC3D to Articulated 3D Body Models

A novel approach for people re-identification is presented in this chapter. The new approach shares some similarities with the previously presented SARC3D algorithm, but extends it to *articulated 3D body models*.

5.1 Introduction

Recently, the introduction of low cost range sensors like the Microsoft Kinect has opened new solutions to surveillance and forensics research fields, included the re-identification task. Thanks to these new devices and the high quality of the estimated depth data it has become possible to detect and track human body joints in real-time [185, 193], this allows for the efficient and effective recognition of human poses and actions. A first attempt at using depth-data for re-identification was made by Barbosa et al. [26], they exploit soft-biometric measures (i.e. set of ratios of joint distances) extracted from the tracked skeletons of pedestrians which are used to generate a person profile. The approach is quite promising, however the intrinsic noise on the estimation of the joint positions do not allow reaching very high performances. For this reason a new approach similar to SARC3D is proposed in this chapter: from the human body joints extracted by a Kinect device, a human skeleton is constructed. The skeleton is sub-divided into

bone fragments, and an appearance-based descriptor is assigned to each one of them.

The high dimensionality of the signatures obtained with this model based approach requires additional steps of feature selection and metric learning. A metric learning approach learns a distance matrix M from a training set of feature vectors which select the most important dimensions and assign to them a higher weight. Different learning algorithms for the metric estimation have been proposed (the most famous being [218]) and adopted in several recent proposals for re-identification [14, 139]. The most important drawback of these methods is the need of a wide training set during the metric learning, in order to avoid the estimation of a metric over-fitted on the specific training data. In this thesis the approach presented by Hirzer et al. [104] is exploited, this method proved to be very efficient and effective for 2D re-identification.

5.2 Bone feature set and person signature

In order to provide a spatial location to each appearance feature, pixels of the color image need to be connected to a bone of the person. To this aim, OpenNi is exploited to find a set of human joints (see the red dots in Figure 5.1), these joints are then linked to build a simple human skeleton (see Figure 5.1). Let $\mathcal{T} = \{\tau_1, \ldots, \tau_{15}\}$ be the set of 15 joints, where each of the elements τ_i corresponds to the 3D position of the joint. Based on $\mathcal{T}$, the corresponding "bone" set $\mathcal{B} \subset \mathcal{T} \times \mathcal{T}$ is obtained as the set of edges of the joint graph (see Figure 5.1). Each bone $\beta_i \in \mathcal{B}$ is defined using the 3D coordinates of the two extremities: $\beta_i = (\tau_r, \tau_s)$. Let $N_\beta = |\mathcal{B}|$ be the number of bones, 18 in our skeleton model.

Given the point cloud $\mathcal{W} = \{(x_1, \kappa_1) \ldots (x_W, \kappa_W)\}$, where x_j and κ_j are the position and the color of each point respectively, a point-to-bone assignment is provided using a min-distance criteria and the subsets $\mathcal{W}_i$ of points connected to the i-th bone β_i are obtained as follows:

$$\mathcal{W}_i = \{(x_j, \kappa_j) \in \mathcal{W} | i = \mathcal{P}(x_j)\}, j = 1 \ldots W \tag{5.1}$$

where $\mathcal{P}$ is the function returning the index of the closest bone:

$$\mathcal{P}(x_j) = \underset{i=1...N_\beta}{\arg\min}\, d(x_j, \beta_i) \tag{5.2}$$

The pixel-to-bone distance $d(x_j, \beta_i)$ is the common point-to-segment Euclidean distance.

After the pixel-to-bone assignment, the signature of the person is composed by the set of color histograms computed for each bone:

$$\mathbf{H}^p = \{H_1^p, \ldots, H_N^p\}\,. \tag{5.3}$$

where H_i^p is the color histogram of the bone β_i of the p-th person. In the experiments reported in chapter 6, H_i^p are RGB color histograms with a 8 bin quantization for each channel, normalized to sum up to 1. If the person model is obtained as the integration of multiple views, the histograms are computed

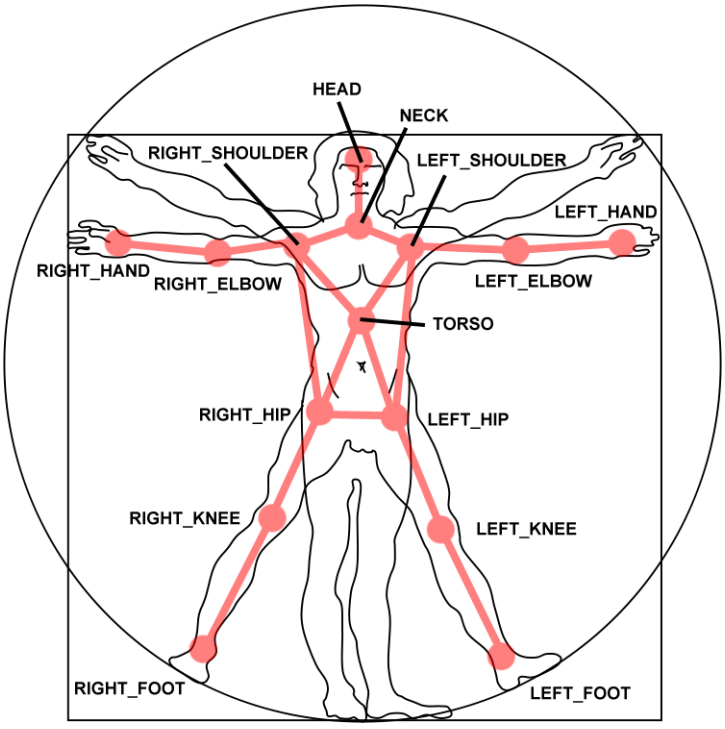
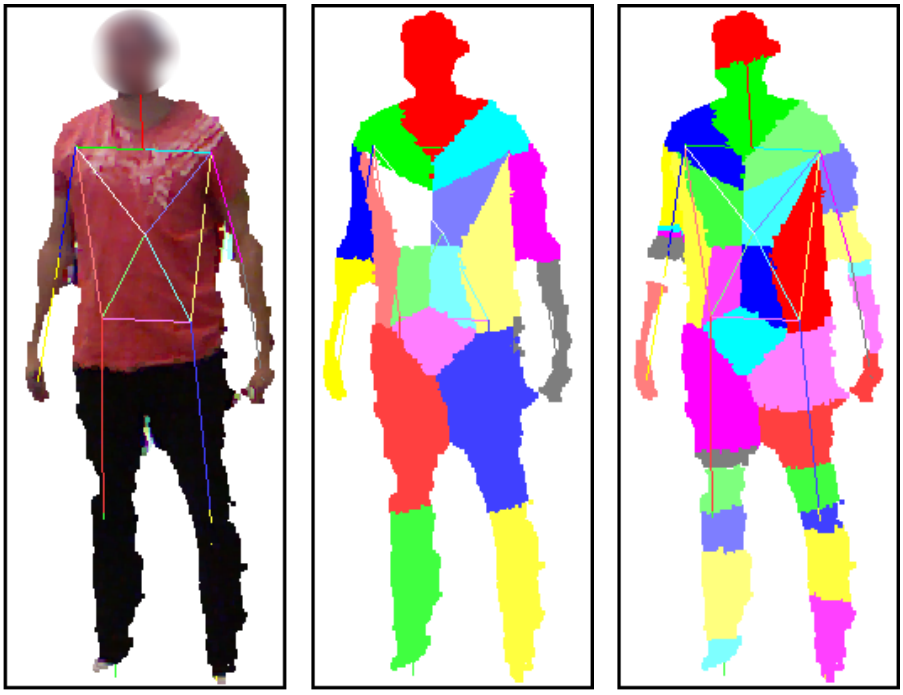


Figure 5.1: Vitruvian body model with superimposed the joints and bones used in the proposed system



(a) (b) (c)

Figure 5.2: Point cloud to bone assignment. a) the skeleton tracking result using the OpenNi libraries, b) the point assignment using the default bone set and c) the final set after automatic fragmentation.

using all the image points assigned to the same bone. A visual example of the pixel-to-bone assignment is reported in Figure 5.2(b).

5.3 Distance metric

The distance between two person signatures $\mathbf{H}^i$ and $\mathbf{H}^j$ can be computed as the sum of distances between each corresponding couple of histograms:

$$d(\mathbf{H}^i, \mathbf{H}^j) = \sum_{n=1}^{N_\beta} \alpha_n \cdot d(H_n^i, H_n^j), \tag{5.4}$$

where $d(H_n^i, H_n^j)$ is the Mahalanobis distance between H_n^i and H_n^j.

However, the distance function of Eq. 5.4 requires the estimation of the bones weights α_n. Moreover, the dimensionality of the signature $\mathbf{H}^p$ computed

for each person as in Eq. 5.3 is too high, leading to problems on its storage and matching. Using the defined skeleton model composed by 18 bones and using 512 bins for each histogram, the final signature is composed by 9216 elements. Each signature is thus processed with a PCA step, which simultaneously reduce the dimensionality and filter the intrinsic noise. The person signature obtained from $\mathbf{H}^p$ becomes a 96 dimensional feature vector $\S_p = \mathcal{X}(\mathbf{H}^p)$.

Instead of using Eq. 5.4, the metric scheme adopted to compare person signatures is inspired from the one presented in [104], called Relaxed Pairwise Metric Learning (RPML), that has proved to be a highly efficient and effective metric learning approach. RPML aims at computing a pseudo-metric M, which, similarly to the Mahalanobis distance, provides a dissimilarity score of two feature vectors $\mathbf{H_i}$ and $\mathbf{H_j}$:

$$\begin{aligned} d_M(\mathbf{H}^i, \mathbf{H}^j) &= (\mathbf{H}_i - \mathbf{H}_j)^\top M (\mathbf{H}_i - \mathbf{H}_j) = \\ &= \| L(\mathbf{H}_i - \mathbf{H}_j) \|^2 \end{aligned} \tag{5.5}$$

where $M = L^\top L$.

In order to exploit the discriminative information of the data during the metric learning, the person re-identification problem is redefined as a two-class problem: firstly, samples from the data space are converted to the label agnostic difference space. Secondly, the original class labels are discarded and the samples are rearranged into the *similar* and *different* classes $\mathcal{S}$ and $\mathcal{D}$ (i.e., if two signatures belong to the same person, their difference is labeled as $\mathcal{S}$, otherwise as $\mathcal{D}$).

Starting from the following objective function:

$$\begin{aligned} \mathcal{L}(L) = \; & \tfrac{1}{|\mathcal{S}|} \textstyle\sum_{(i,j)\in\mathcal{S}} ||L(\mathbf{H}_i - \mathbf{H}_j)||^2 \\ & - \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{(i,j)\in\mathcal{D}} ||L(\mathbf{H}_i - \mathbf{H}_j)||^2 \end{aligned} \tag{5.6}$$

the problem of finding the best M can then be defined as

$$\begin{aligned} \arg\min \quad & \mathcal{L}(M) \\ \text{subject to} \quad & M \succeq 0, \\ & L\Sigma_S L^\top = I, \\ & L\Sigma_D L^\top = I \end{aligned} \tag{5.7}$$

where

$$\mathcal{L}(M) = tr(M(\Sigma_S - \Sigma_D)) \tag{5.8}$$

and

$$\Sigma_S = \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} (\mathbf{H}_i - \mathbf{H}_j)(\mathbf{H}_i - \mathbf{H}_j)^\top \tag{5.9}$$

$$\Sigma_D = \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} (\mathbf{H}_i - \mathbf{H}_j)(\mathbf{H}_i - \mathbf{H}_j)^\top \tag{5.10}$$

By relaxing the positivity constraint $M \succeq 0$, the problem can be simplified into finding an M such that

$$tr(M(\Sigma_S - \Sigma_D)) = 0 \tag{5.11}$$

A bounded approximation for M can be found as $M = ({\Sigma_S}^{-1} - {\Sigma_D}^{-1})$. Technically, due to the relaxation the finally obtained matrix M does not describe a pseudo-metric. Nevertheless, the experimental results show that the estimated solution provides a sufficient approximation for the given task and that competitive results can be obtained; however, on a much lower computational effort than other metric learning approaches.

5.4 Automatic bone fragmentation

The bone set $\mathcal{B}$ defined using the fifteen joints from OpenNi is not optimized for people re-identification. The main drawbacks are related to the different lengths of the bones and the semantic differences between bones and appearance regions. To this aim, an automatic bone fragmentation step has been introduced in order to generate a set of bones $\mathcal{B}^*$ more suitable for re-identification tasks. Our goal is to find the skeleton partitioning which maximizes the re-identification score AUC (see Section 3.4.2). The size of the corresponding search space is too high for an exhaustive global maximum selection. An iterative procedure inspired by the Cuckoo search approach [210] has been implemented.

The set $\mathcal{B}^*$ is iteratively derived from $\mathcal{B}$ replacing one of the existing bones with the couple of sub-bones obtained after a split. A split $S = (b_i, \alpha)$ is represented by the index of the selected bone β_i and the split position $\alpha \in [0, 1]$.

At each iteration, a set of candidates $\{S^k\}$ is randomly generated by selecting both the bone index and the split point α using a fast uniform distributed generator. The best split is then selected maximizing the re-identification score AUC on a training set of samples. The iterative algorithm is stopped when the re-identification performance are not increased by adding an additional split. The final bone set created using our training set and the corresponding pixel-to-bone assignment are depicted in Figure 5.2(c) and is composed by 38 elements. An experimental evaluation of the proposed system is presented in chapter 6, section 6.3.

Chapter 6

Experiments

In this chapter experimental results of the proposed methods for people re-identification and their components are presented. The summarized performance results are reported using the Cumulative Matching Characteristic (CMC) curve, which is analogous to the ROC curve for detection problems[96], and using the compact Area Under the Curve (AUC) measure, which is the integral of the cumulative curve. Firstly, to demonstrate the use of the CMC curve and relative measures, experimental results of state of the art algorithms on the ViPER and CAVIAR datasets are reported. Then an extensive experimental evaluation of the proposed methods (SARC3D and 3D articulated body models) is presented. Lastly and in-depth experimental evaluation of the proposed orientation detection algorithm is reported.

6.1 Reported results from State of the Art methods

To demonstrate the use of the CMC curve and relative measures, Table 6.1 and Table 6.2 contain quantitative comparisons of some of the techniques reviewed in Chapter 2 on the ViPER and CAVIAR datasets (using two different subsets), respectively. The numeric values have been collected from the corresponding papers and some of them have been estimated approximately from graphical outputs. Excluding the last three rows obtained with interactive refinements

Table 6.1: Quantitative comparison of some methods on the ViPER dataset

Method	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
RGB Histogram	0,04	0,11	0,20	0,27
ELF [95]	0,08	0,24	0,36	0,52
Shape and Color Covariance Matrix [146]	0,11	0,32	0,48	0,70
Color-SIFT [146]	0,05	0,18	0,32	0,52
SDALF [80]	0,20	0,39	0,49	0,65
Ensemble-RankSVM [40]	0,16	0,38	0,53	0,69
PRDC [218]	0,15	0,38	0,53	0,70
MCC [218]	0,15	0,41	0,57	0,73
IML [6]				
- No Metric Learning	0,07	0,11	0,14	0,21
- using human interaction - 1 iteration	0,42	0,42	0,43	0,50
- using human interaction - 5 iterations	0,74	0,74	0,74	0,74
- using human interaction - 10 iterations	0,81	0,81	0,81	0,81

through relevance feedback, the best Rank-1 performance on the ViPER dataset is provided byFarenzena et al. [80](SDALF). Sophisticated ranking strategies could improve the performance of Rank-5, Rank-10 and Rank-20, as highlighted in the results reported by Brosser et al. [40] and Zheng et al. [218]. A relevance feedback step proved to be very helpful with retrieving all the required matches. This could be a useful application in forensics but less so in automatic surveillance applications where human interaction is uncommon.

6.2 SARC3D Experimental results

In order to evaluate the SARC3D method presented in Chapter 4 several tests were performed on these publicly available datasets:

- 3DPeS[24] (See section 3.3).
- the ViSOR Dataset[23] (See section 3.2).

Table 6.2: Quantitative comparison of some methods on selected frames from the CAVIAR dataset

Method	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
Caviar4Reid data set				
AHPE[29]	0,08	0,32	0,53	0,74
SDALF [80]	0,09	0,38	0,58	0,77
CPS [50]	0,16	0,48	0,69	0,86
Personal data set				
BoF+SVM [135]	0,58	0,93	0,98	1,00

- the PETS2009 dataset[166].

3DPeS was used for testing the performances of the full SARC3D proposal, while the ViSOR dataset was exploited to test the SARC3D model resilience to errors and noise in the orientation and position estimation. Some qualitative outcomes on the PETS2009 dataset are shown. Firstly different combinations of features, metrics and parameters are tested. Then, a comparison with two state of the art methods is provided. Each SARC3D experiment is shown using CMC curves computed averaging 10 different test runs on random partitions of the dataset. Lastly the SARC3D model resilience to errors and noise in the orientation and position estimation is evaluated. Some of the experiments reported here were also partially presented in [23]

6.2.1 SARC3D Feature selection and parameter tuning

The performances of the vertex descriptors proposed in section 4.2.1 and the different strategies for multiple views integration of section 4.4 are reported: Mean HSV histogram, Mean RGB histogram, Bag of HSV histograms, Bag of RGB histograms and a combination of RGB and HoG histograms (both in the variants bag-of-histograms and average histograms) are some of the vertex descriptors tested. All tests were conducted on the 3DPeS dataset.

For this particular test 3 views for the creation of both the query and the set of training models were randomly selected. Figure 6.1(a) shows the corresponding CMC curves. The bag-of-histograms solutions always outperform the averaged alternatives, demonstrating stronger reliability to viewpoint changes and alignment

	HSV Average	HSV BoH	RGB Average	RGB BoH	RGB+HoG Average	RGB+HoG BoH
3vs1	0.889	0.914	0.882	0.905	0.872	0.898
3vs3	0.961	0.971	0.967	0.975	0.968	0.977

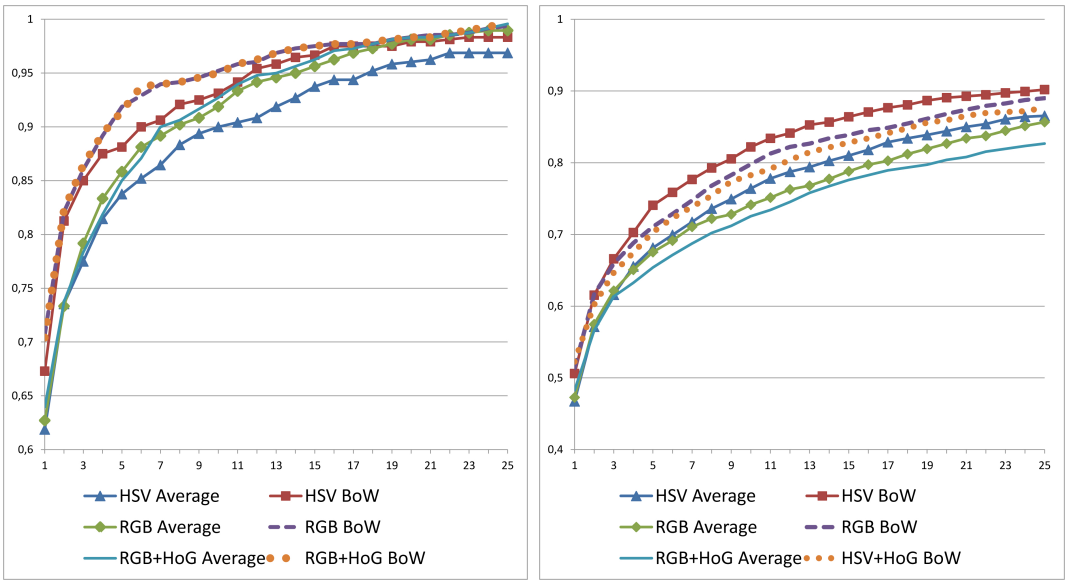
Table 6.3: AUC for the different features tested: 3vs3 and 3vs1 tests

errors.

To test the performance of the model in unbalanced cases, i.e., when the matching models are generated by a different number of views, a different training set was created, using 3 views as in the previous experiments, while the query model was created from a single view only (indicated as *3vs1* in the following). Figure 6.1(b) shows the corresponding results.

Table 6.3 shows the AUC (Area Under Curve) measure for the six types of feature vectors, the first rows shows tests performed using 3 views for both the creation of the query and test models (*3vs3*), the second row report tests performed using 3 views for the query model creation and only 1 view for the test model (*3vs1*). As can be seen from the table the HoG feature barely improves performance, this is due to the presence of many low-resolution images in the dataset. If the dataset is restricted to high-resolution images only the HoG features improve performance by $1-5\%$. As an example, when restricting the dataset to images with a resolution greater than $150x50$ pixels the resulting AUCs for the *3vs1* and *3vs3* test cases are 0.971 and 0.983 for the RGB+HoG feature and 0.954 and 0.976 for the RGB only feature, both in the Bag of Histograms variant.

Finally, the effective capability of SARC3D to integrate multiple views was evaluated. In figure 6.2 the improvement in re-identification performances obtained by adding more views to each SARC3D model are reported. The results are also reported in Table 6.4 for all the tested cases using the AUC score. Figure 6.2 also compares HSV and RGB features in the Bag-of-Histograms configuration: as can be inferred from the graphs and from table 6.4, HSV outperform RGB when the number of views used is low. This may be due to the HSV space being slightly more resilient to illumination and color response changes, but when lots of additional views are available the higher color discriminability of the RGB space outperform HSV, despite being more sensible to noise (like illumination and color



(a) (b)

Figure 6.1: Testa on different feature sets: a) query and test models created with 3 views each, b) query models created with 3 views, test models created with one view only

changes) a problem which is however partially solved by the bag-of-histograms approach.

	1vs1	2vs1	3vs1	4vs1	5vs1	3vs3
Median RGB	0.812	0.852	0.882	0.942	0.958	0.967
Median HSV	0.827	0.868	0.883	0.938	0.958	0.961
HSV+BoH	0.830	0.872	0.914	0.950	0.963	0.971
RGB+BoH	0.817	0.856	0.905	0.954	0.967	0.975

Table 6.4: AUC for all the tested cases and different number of views

In the next series of experiments, a particular descriptor was selected (average HSV histograms) and the resilience of the SARC3D model to random perturbation of it's position and orientation was tested. All tests were performed on the ViSOR dataset. In tables 6.5 and 6.6 the system performance in presence of random perturbations of the correct alignment is reported; both errors on the ground plane localization and on the orientation have been introduced. The performance reported on the table shows that our system is still reliable, even in the case of

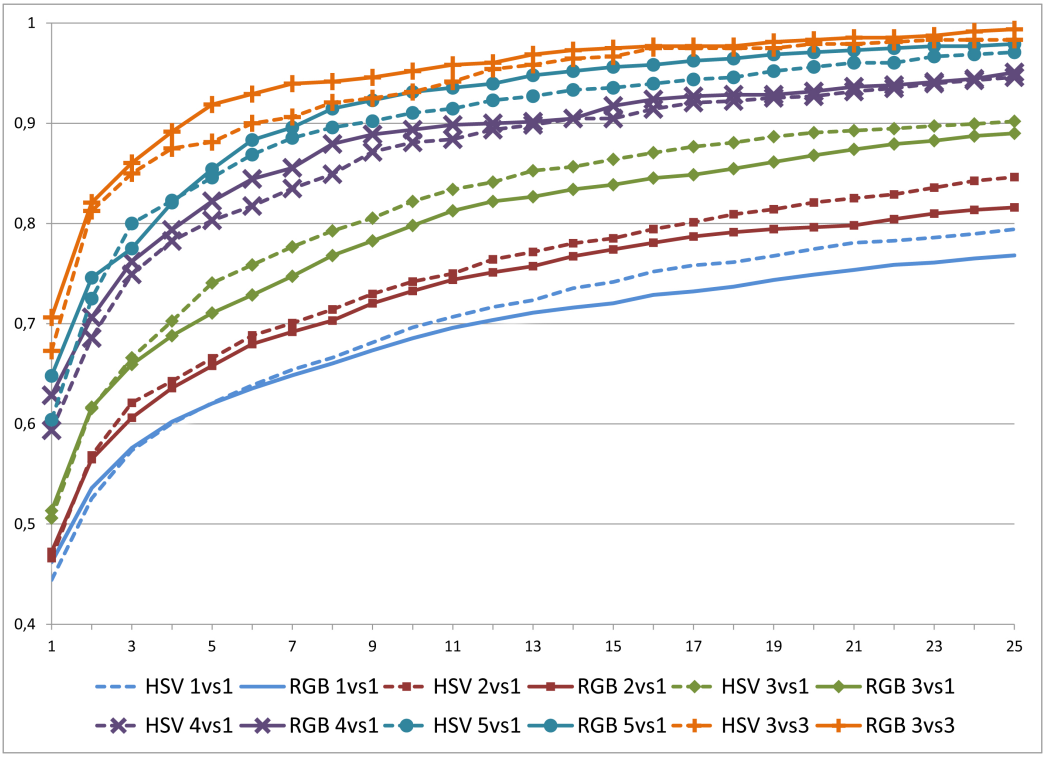


Figure 6.2: Multiview comparison between HSV and RGB features in the Bag-of-Histograms configuration

non-precise model alignment and orientation, keeping good results with localization precision up to 6 pixels (45% overlap between the projected bounding box of the model and the image blob) and orientation up to 30 degrees.

6.2.2 SARC3D Comparison with state of the art

In this section SARC3D is compared with two state-of-the-art methods on the 3DPeS dataset: the SDALF method proposed by Farenzena et al. [80] and the ensemble of features proposed by Gray and Tao [95].

SDALF is a purely two dimensional method. It consists in the extraction of features that model three complementary aspects of the human appearance: the overall chromatic content (using weighted HSV histograms), the spatial arrangement of colors into stable regions (Maximally Stable Color Regions), and the presence of recurrent local motifs with high entropy. All these features are

Rank	Correct Alignment	With noise on localization				
		2 px (15%)	4 px (30%)	6 px (45%)	12 px (75%)	16 px (90%)
1	0.68	0.63	0.60	0.58	0.51	0.45
2	0.78	0.75	0.73	0.72	0.65	0.60
3	0.83	0.83	0.81	0.80	0.74	0.68
4	0.85	0.86	0.85	0.84	0.76	0.75
5	0.90	0.88	0.86	0.85	0.78	0.78
10	0.96	0.96	0.97	0.95	0.94	0.90

Table 6.5: Performance evaluation of the system using random perturbations of the 3D model localization (3vs1 case)

Rank	Correct Alignment	With noise on orientation						
		5°	10°	15°	30°	40°	60°	90°
1	0.68	0.61	0.60	0.58	0.55	0.54	0.53	0.50
2	0.78	0.74	0.75	0.76	0.70	0.69	0.69	0.66
3	0.83	0.80	0.81	0.81	0.77	0.75	0.81	0.75
4	0.85	0.85	0.84	0.83	0.81	0.79	0.85	0.80
5	0.90	0.89	0.87	0.88	0.86	0.82	0.87	0.84
10	0.96	0.96	0.96	0.96	0.94	0.95	0.96	0.95

Table 6.6: Performance evaluation of the system using random perturbations of the 3D model orientation (3vs1 case)

derived from different body parts, and opportunely weighted by exploiting symmetry and asymmetry perceptual principles (each appearance image is segmented into legs/torso/head using simple heuristics).

The method proposed in [95] consists of an ensemble of features: RGB, HSV and YCbCr histograms (each channel quantized into 16 bins) and the histograms of the response to 13 Schmid and 8 Gabor filters (each response quantized into 16 bins). The different features are concatenated in a single 464 dimensional feature vector, a feature vector is computed for each of the 3 fixed size stripes of the person silhouette (roughly head, torso and legs). This method is extended to the multi-view case by exploiting a bag-of-feature approach. In order to have a fair comparison, the distance between feature vectors is computed through the Euclidean distance between the descriptors instead of applying an additional metric learning.

Figures 6.3 and 6.4 shows different comparisons between the aforementioned methods. Specifically figure 6.3(a) shows the case when 1 image only is used for

both the training and the query model (indicated as *1vs1* in the following), 6.3(b) when 3 images are used for the training model (indicated as *3vs1* in the following), 6.4(a) when 5 images are used for the training model (indicated as *5vs1* in the following) and 6.4(b) when 3 images are used for both the training and the query model (indicated as *3vs3* in the following). Table 6.7 reports accuracies at ranks 1, 5, 10 and 25 for the aforementioned four graphs.

As it can be inferred from the presented results, multi-views approach always outperform single shot approaches, however, adding 3D information, and specifically a better feature localization scheme like SARC3D, greatly improves the re-identification performances as highlighted by table 6.7. It also should be noted that the presented algorithm does not require a pixel perfect position and orientation estimation. As can be inferred from figures 6.20 and 3.4 the position and orientation estimation is far from perfect (reaching 61% accuracy on the 3DPeS dataset as reported in the following section), at the same time the silhouettes extracted and exploited in our methods are usually very noisy with lots of holes and missing parts as shown in fig.3.4 where some sample silhouettes are reported. Nevertheless, SARC3D substantially outperforms the other methods at every rank.

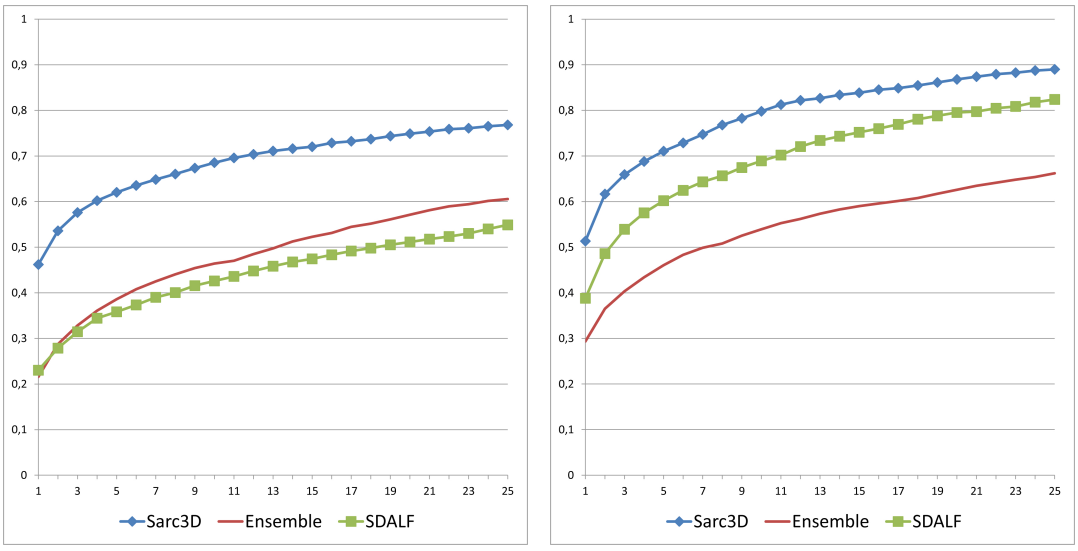
6.2.3 SARC3D Qualitative experiments

The system was also tested on the PETS 2009 dataset [166] in order to evaluate the proposed method in real life conditions. The *City center* sequence, with three overlapping camera views, was selected. A 12.2m $\times$ 14.9m ROI was chosen, which is visible from all cameras. The proposed method was added on top of a previously developed tracking system (See Appendix A for details), the goal of our method was to repair broken track and re-identify people that enter and exit the rectangular ROI. Fig. 6.5 shows some sample frames from the system in action. The obtained precision and recall are 80.2% and 88.7% respectively.

Fig. 6.6 shows three particular sample results from the ViSOR dataset: the SARC3D model allows to evaluate small features in the people appearance and it's able to discriminate between three people wearing extremely similar dresses, which differs only in minute details, like neck shirt shape and other small peculiar

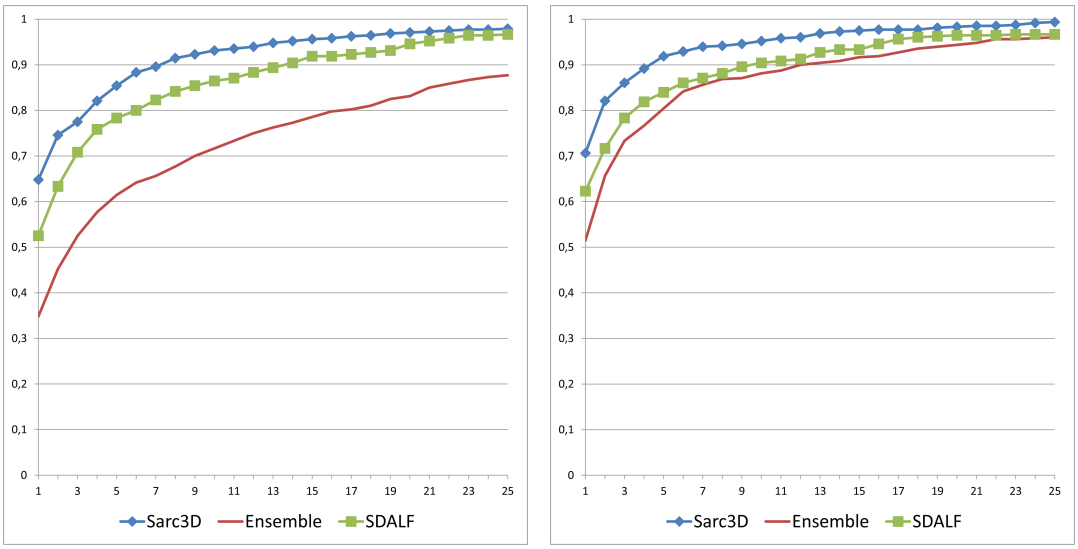
View integration	Rank	SARC3D	SDALF	Ensemble
1vs1	1	46%	23%	21%
	5	62%	35%	38%
	10	68%	42%	46%
	25	76%	54%	60%
3vs1	1	51%	38%	29%
	5	71%	60%	46%
	10	79%	68%	53%
	25	89%	82%	66%
5vs1	1	64%	52%	34%
	5	85%	78%	61%
	10	93%	86%	71%
	25	97%	96%	87%
3vs3	1	70%	62%	51%
	5	91%	83%	80%
	10	95%	90%	88%
	25	99%	96%	96%

Table 6.7: Average accuracy at ranks 1, 5, 10 and 25 using RGB histograms with BoH for both the single shot and the multi-shot cases



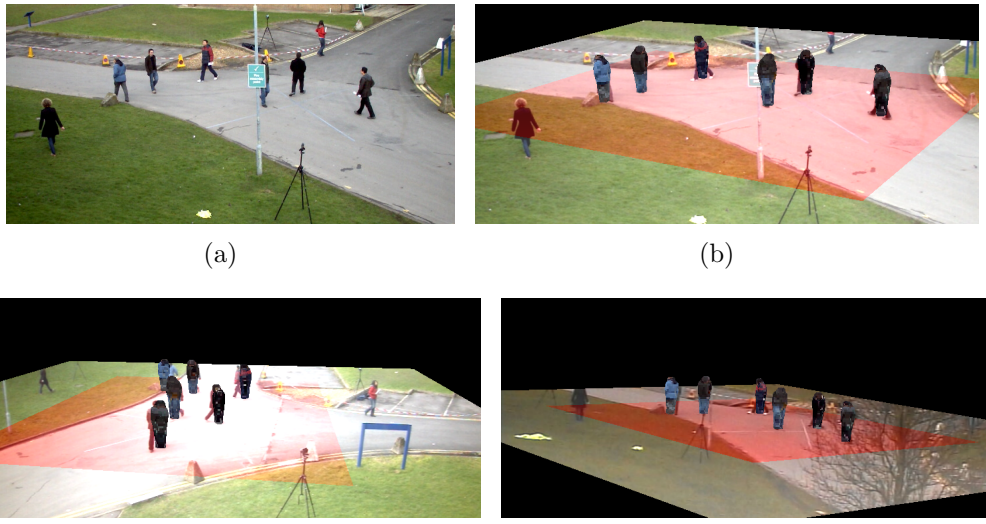
(a) (b)

Figure 6.3: Comparisons with the state of the art: a)single shot - *1vs1*, b) multi-shot *3vs1*



(a) (b)

Figure 6.4: Comparisons with the state of the art: a) multi-shot *5vs1*, b) multi-shot *3vs3*

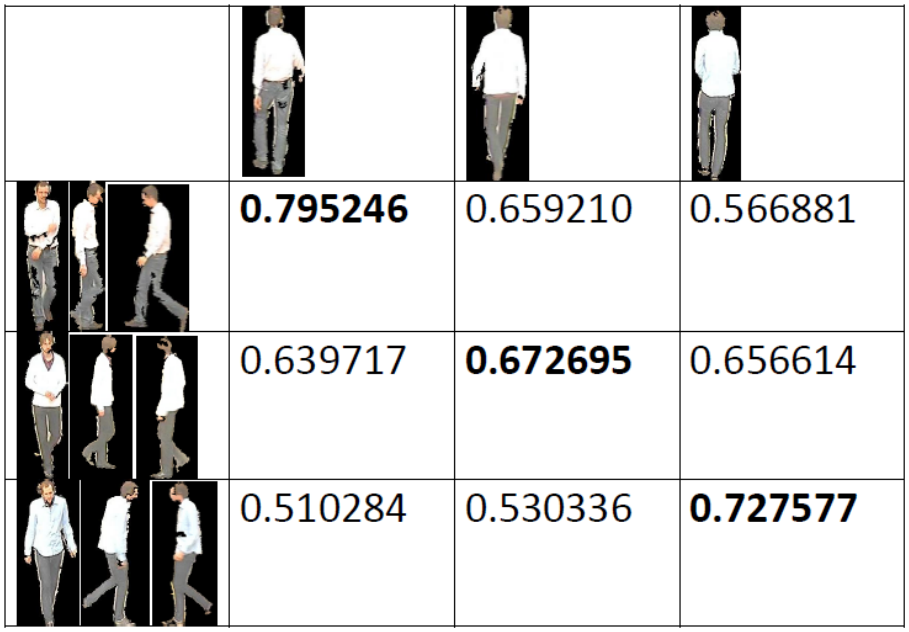


(a) (b)

(c) (d)

Figure 6.5: (a) PETS dataset, sample frame from camera 1, (b,c,d) system output superimposed to camera 1, 2 and 3 frames

aspects of the shirts design. Figure 6.7 shows some selected test queries made on the 3DPeS dataset, comparing SARC3D and SDALF returns.



(a)

Figure 6.6: Distance matrix obtained from three very similar people: the three images used for the model creation (rows) and the test images (columns) are also shown.

6.3 3D Articulated Body Model Experimental evaluation

In order to evaluate the method proposed in Chapter 5 a new dataset was created using Microsoft Kinect to extract depth information and relative point cloud. The dataset contains various images of 40 people in different poses and orientations, for a total of 450 shots and relative skeleton and point clouds. Half of the images were used for metric learning and as training set for the bone fragmentation learning, while the remaining 225 shots as testing. Some sample images from the dataset are reported in Figure 6.8.

In order to evaluate the contribution of the bone fragmentation algorithm described in Section 5.4 and the metric learning defined in Section 5.3 internal comparisons were initially performed. The results obtained using the base system (i.e., using the OpenNi bone set and the distance function of Eq. 5.4 with

(a) (b)

Figure 6.7: Example queries made to our re-identification database. (a) Probe image (for SARC3D this is just one of the images used for the model creation). (b) Top 10 results (sorted left to right). First row shows SDALF results, second row SARC3D results. The correct match is highlighted in green.

uniform weights) and with the integration of the two learning steps are reported in Figure 6.9. The curves have been obtained averaging 50 experiments with different training and the testing sets (randomly selected). As highlighted by the graph, the learned metric strongly improves the re-identification performance. The corresponding AUC values are reported in Table 6.8.

The proposed method was also compared to two state of the art re-identification algorithms, SDALF [80] and SARC3D (detailed in Chapter 4). The online code has been adopted and applied to the RGB images after the removal of the background by means of the depth stream. The head and feet positions from the skeleton stream have been adopted as reference to align the SARC3D model on the images. Results of the three methods are reported in Figure 6.10. The pro-

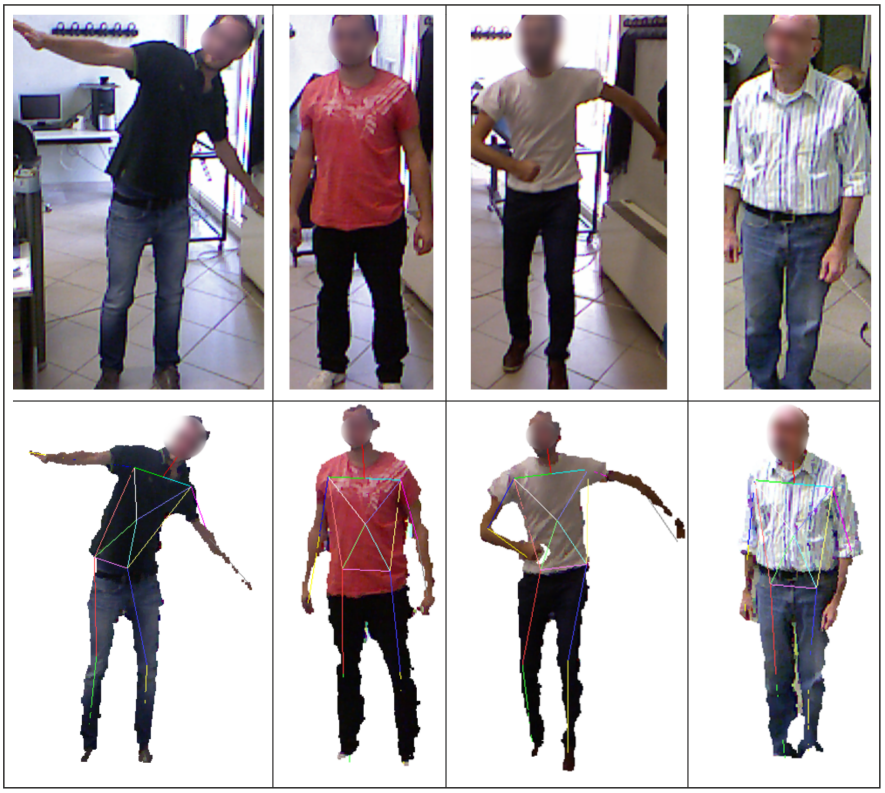


Figure 6.8: Sample images and corresponding point clouds with the estimated skeletons from the Kinect dataset

Method	AUC
Base	0.963
+ Auto. bone fragm.	0.971
+ Metric learning	0.990
Overall proposal	**0.994**
Sarc3D [23]	0.983
SDALF [80]	0.969

Table 6.8: AUC values of the methods tested

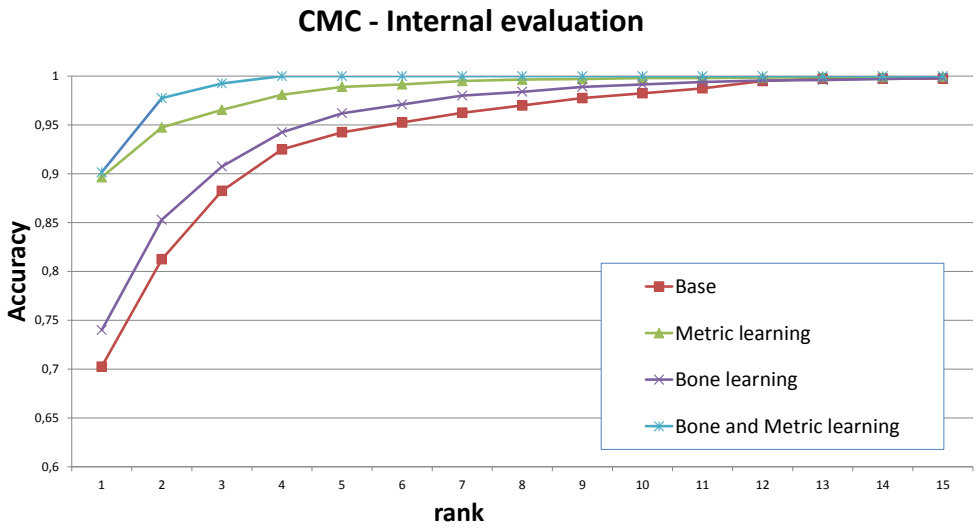


Figure 6.9: CMC curves of the proposed method on the Kinect dataset, showing the contribution of the metric and bone fragmentation learning

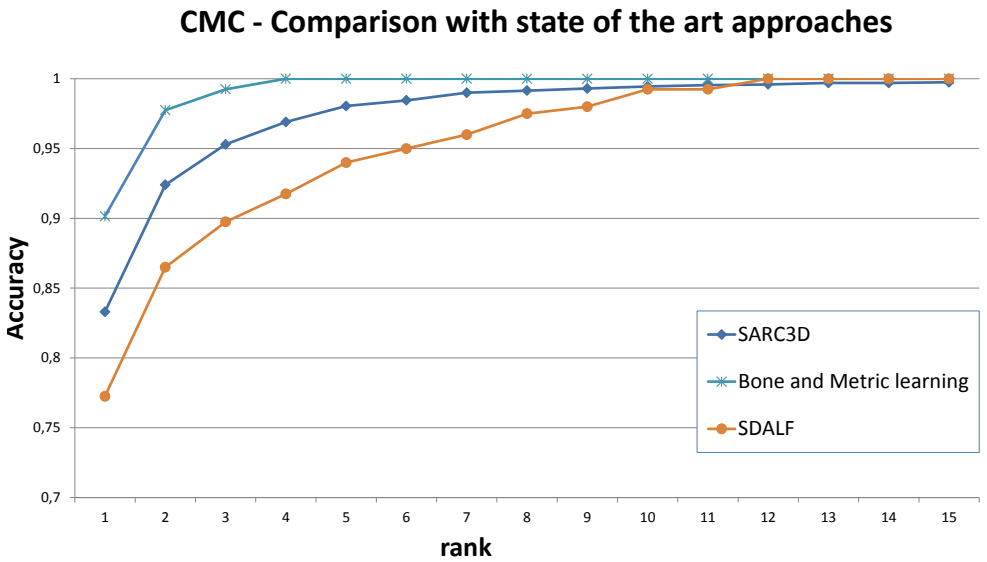


Figure 6.10: CMC curves of the proposed method (with both metric and bone fragmentation learning) and two state of the art techniques on the Kinect dataset

posed method outperforms both the SARC3D and the SDALF approaches. The proposed method requires an off-line training in order to learn the metric (which takes on average 4 seconds) and to learn the best splits of the skeleton (which takes on average 195 minutes). The creation of a person model requires less than 10 ms (8.74ms for the computation of the point cloud by OpenNI and 0.47ms for the computation of the histograms), the matching score computation takes on average 2.11 ms. All tests were performed on an Intel Core i5 running at 2.66 GHz.

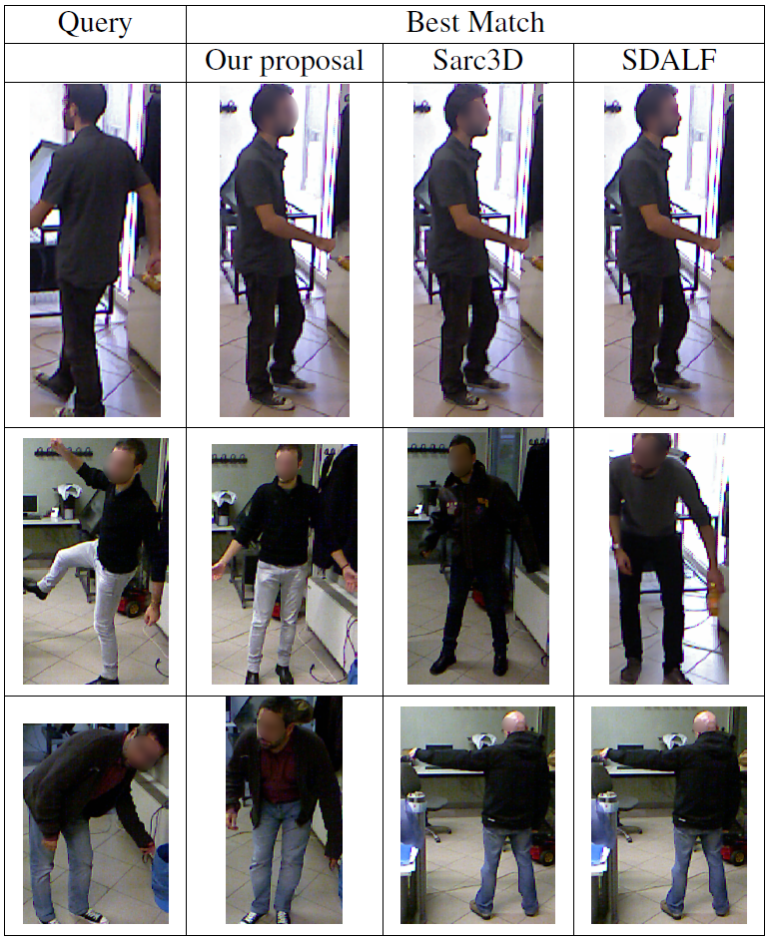


Figure 6.11: Sample results of the three tested methods on our dataset

Some sample results of the three methods on the new dataset are shown in Figure 6.11, as can be seen the proposed method is much more resilient to pose changes, unlike SDALF and SARC3D which assume people in a vertical standing

position.

6.4 Orientation Detection Experimental results

The proposed Orientation Detection approach has been extensively tested on three public datasets, namely the TUD Multiview Pedestrians dataset [11], 3DPeS [24] and The ViSOR dataset [23]. In the following, quantitative results are presented and compared to the state of the art; finally some qualitative outcomes on video sequences from PETS and 3DPeS are shown. The experimental results reported here were also presented in [25].

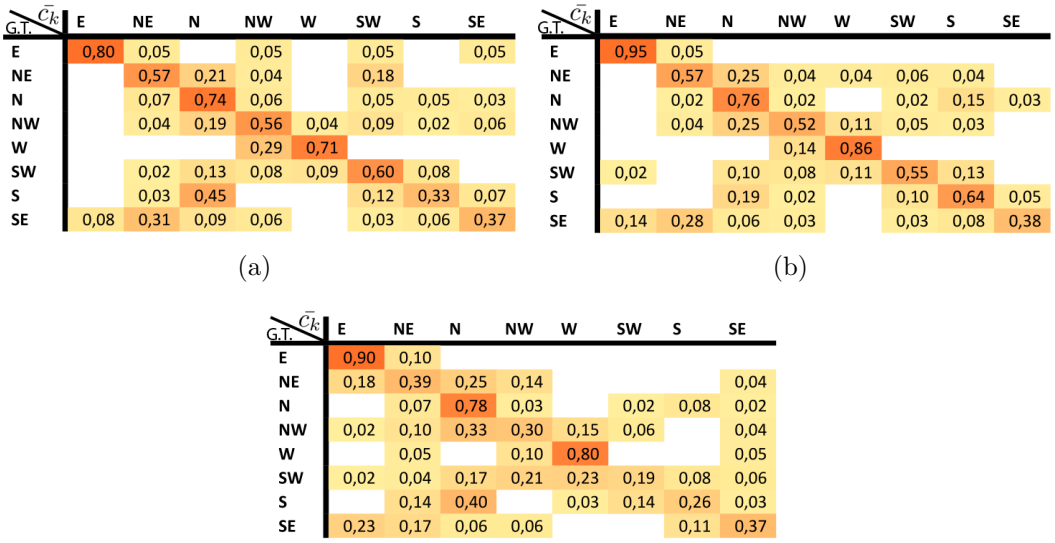
The TUD Multiview Pedestrian dataset contains 5288 snapshot of pedestrians, fully annotated with bounding boxes, orientation classes and skeletons. 20% of the images were randomly selected from the provided training set to train the classifiers, and then the same split originally proposed for validation and testing are use: 248 snapshots for validation and parameters estimation, and 300 images for testing.

For the 3DPeS dataset, people snapshots were randomly selected from the provided videos and manually annotated, obtaining 1012 snapshots, 360 used for training, 652 for testing.

The ViSOR Dataset provides 200 snapshots of 50 people taken from 4 predefined points of view. All the provided images were used for testing only, exploiting the same classifiers and parameters learned from 3DPeS, since images have similar characteristics (image resolution and camera point of view).

Test results are presented in terms of classification confusion matrices, where each row contains the ground-truth label whilst each column indicates the predicted one. In the following it is reported the classification accuracy for each class, and two final measures: "Accuracy 8", where exact hits only are considered, and "Accuracy 4" where adjacent classes are also considered correct.

The first extensive experiments were carried out on the TUD Multiview Pedestrian dataset. Fig. 6.12 shows the results of our method without (Fig. 6.12(a)) and with (Fig. 6.12(b)) the Mixture of Approximated Wrapped Gaussians filtering step. Predictably, the intermediate directions are more difficult to recognize; additionally opposite and specular directions are difficult to disambiguate.



(a)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,80	0,05		0,05		0,05		0,05
NE		0,57	0,21	0,04		0,18		
N		0,07	0,74	0,06		0,05	0,05	0,03
NW		0,04	0,19	0,56	0,04	0,09	0,02	0,06
W				0,29	0,71			
SW		0,02	0,13	0,08	0,09	0,60	0,08	
S		0,03	0,45			0,12	0,33	0,07
SE	0,08	0,31	0,09	0,06		0,03	0,06	0,37

(b)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,95	0,05						
NE		0,57	0,25	0,04	0,04	0,06	0,04	
N		0,02	0,76	0,02		0,02	0,15	0,03
NW		0,04	0,25	0,52	0,11	0,05	0,03	
W				0,14	0,86			
SW	0,02		0,10	0,08	0,11	0,55	0,13	
S			0,19	0,02		0,10	0,64	0,05
SE	0,14	0,28	0,06	0,03		0,03	0,08	0,38

(c)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,90	0,10						
NE	0,18	0,39	0,25	0,14				0,04
N		0,07	0,78	0,03		0,02	0,08	0,02
NW	0,02	0,10	0,33	0,30	0,15	0,06		0,04
W		0,05		0,10	0,80			0,05
SW	0,02	0,04	0,17	0,21	0,23	0,19	0,08	0,06
S		0,14	0,40		0,03	0,14	0,26	0,03
SE	0,23	0,17	0,06	0,06			0,11	0,37

Figure 6.12: Confusion matrices on the TUD Multiview Pedestrian Dataset: (a) without MoAWG filtering, (b) with MoAWG filtering, (c) using only the 4 main classifiers (E, N, W, S) and the MoAWG step. Each row contains the ground-truth label whilst each column indicates the predicted one

Directly using the outputs of the classifiers the average accuracy is around 58%. However, the MoAWG step usually improves the classification of ambiguous cases, reaching an average overall accuracy of 65%. In order to highlight the contribution of the MoAWG step, an additional test was performed: the training set was reduced to just four main directions instead of eight, generating an array of 4 classifiers only (E, N, W, S). The continuous p.d.f. of Equation 4.17 is obtained as a Mixture of four components, but the final label is quantized in 8 classes, recovering the intermediate directions. The corresponding confusion matrix is reported in fig. 6.12(c).

Fig. 6.13(a) shows the confusion matrix of the proposed method on the 3DPeS dataset, while Fig. 6.13(b) shows the system results on the ViSOR dataset. The average accuracy of the system is still around the 60% in both cases. The ViSOR dataset contains people oriented in 4 main directions only, but the array of 8 classifiers trained on the 3DPeS dataset was exploited.

Fig. 6.14 summarizes the results obtained on the three datasets, both with the

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,86	0,12						0,02
NE	0,31	0,60	0,07			0,01		0,01
N	0,02	0,05	0,69	0,15	0,01	0,01	0,06	0,01
NW			0,08	0,51	0,29	0,04	0,04	0,04
W	0,15			0,15	0,70			
SW	0,04	0,01		0,02	0,29	0,56	0,08	
S	0,01	0,02	0,22	0,04		0,11	0,47	0,13
SE	0,32		0,05	0,05			0,05	0,53

(a)

G.T. \ $\bar{C}_k$	E	NE	N	NW	W	SW	S	SE
E	0,52	0,46				0,02		
NE								
N		0,24	0,74				0,02	
NW								
W	0,02			0,06	0,44	0,44	0,04	
SW								
S		0,02	0,26			0,06	0,66	
SE								

(b)

Figure 6.13: Confusion matrices on the (a) 3DPeS and (b) Sarc3D datasets.

	E	NE	N	NW	W	SW	S	SE	Accuracy 8	Accuracy 4
TUD	0,95	0,57	0,76	0,52	0,86	0,55	0,64	0,38	0,65	0,83
3DPeS	0,86	0,6	0,69	0,51	0,7	0,56	0,47	0,53	0,61	0,89
Sarc3D	0,52		0,74		0,44		0,66		0,59	0,87
TUD - No AWG	0,8	0,57	0,74	0,56	0,71	0,6	0,33	0,37	0,58	0,76
3DPeS - No AWG	0,82	0,68	0,69	0,66	0,54	0,58	0,38	0,42	0,59	0,87
Sarc3D - No AWG	0,3		0,78		0,26		0,54		0,47	0,9

Figure 6.14: Performance summary on the three datasets

MoAWG step and without it (indicated as No AWG in the table). As reported, the MoAWG step always improves the classification performance (from 4% on 3DPeS, to 25% on Sarc3D). Additionally accuracy 8 and accuracy 4 scores are reported. A visual example of the classification output on the TUD dataset is depicted in Fig. 6.15. Each image has been recolored following the rules of Fig. 4.14. The top half color of each image represents the ground-truth value, while the bottom half shows our results.

Using the same method, qualitative results on an excerpt of the "PETS2009 - Flow Analysis and Event Recognition" video sequence and on a video from 3DPeS are shown in Fig. 6.16 and Fig. 6.17 respectively. In this last case, the ground-truth was generated from the person trajectory on the ground plane and thus a precise orientation angle is used in the evaluation.

6.4.1 Comparative evaluation

The system results have been compared against two state of the art techniques [11, 48] and other alternative solutions exploiting different classifiers and features on the TUD dataset. In particular, the classification accuracy of the system is

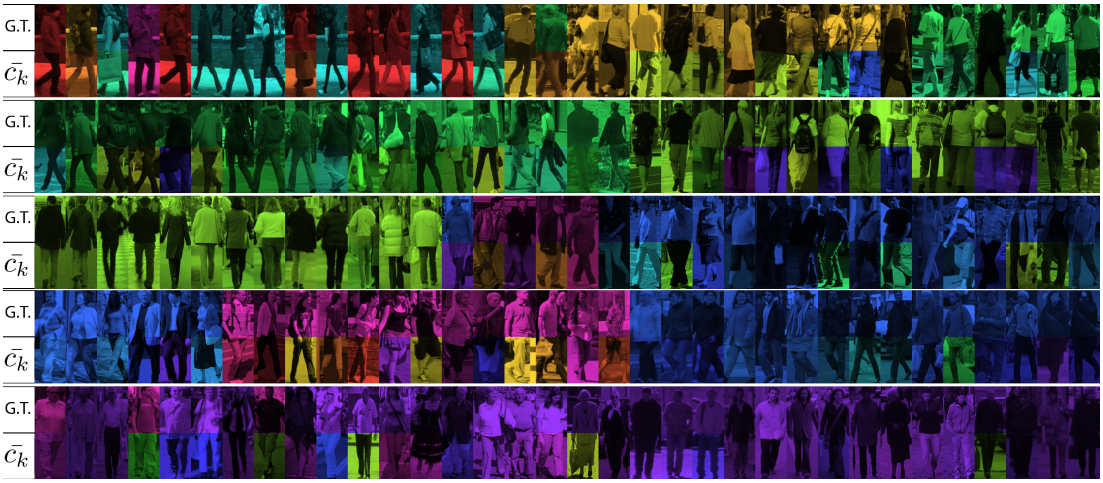


Figure 6.15: Qualitative results on some snapshots from the TUD dataset

Figure 6.16: Qualitative results on a excerpt from PETS2009; a woman dressed in red is initially walking from left to right and then away from the camera

evaluated, where variations of Support Vector Machines and Random Forest classifiers are adopted instead of the Extremely Randomized Trees, and Covariance Descriptors [196] replace the HoG-based feature vector (Table 6.18). The first row (HoG - ERT -AWG) reports the performances of our complete method which outperforms all the other solutions:

- Eight Randomized Forests on HoG features with and without the MoAWG step (HoG - RT - AWG and HoG - RT - NoAWG);
- Eight SVMs in regression mode on HoG features with and without the MoAWG step (HoG - Multi SVMr - AWG and HoG - Multi SVMr - NoAWG);
- A single multi-class SVM used in Classification mode on HoG features (HoG - Single SVMC - No AWG); the response vector ψ is not available and thus

Figure 6.17: Qualitative results on images from a person tracked in 3DPeS

the AWG step is not applicable;

- A single Multi-class Random Trees classifier on HoG features (HoG - Single RT - No AWG); similarly to the SVM based solution, the AWG step is not allowed in this case;

- Just four Extremely Randomized Forests on HoG features with and without the MoAWG step (HoG - 4 ERT - AWG and HoG - 4 ERT - NoAWG); The AWG step allows to recover the intermediate directions.

- Extremely Randomized Trees on classic Covariance descriptors [196] (COV - ERT - AWG and COV - ERT - NoAWG);

- Eight SVMs in regression mode on classic Covariance descriptors [196] (COV - SVM - AWG and COV - SVM - NoAWG);

- Extremely Randomized Trees and SVMr on modified Covariance descriptors, which embed information on mutual correlation between gradients in different image cells (COV2 - ERT - AWG, COV2 - ERT - NoAWG, COV2 - SVM - AWG and COV2 - SVM - NoAWG);

In all experiments the following parameters where used: for the Extremely Randomized Trees and Random Forests Classifiers the number of trees was set

	E	NE	N	NW	W	SW	S	SE	Overall
HoG - ERT - AWG	0,95	0,57	0,76	0,52	0,86	0,55	0,64	0,36	0,65
HoG - ERT - NoAWG	0,8	0,57	0,74	0,56	0,71	0,6	0,33	0,37	0,58
HoG - RT - AWG	0,73	0,57	0,58	0,45	0,55	0,68	0,5	0,32	0,54
HoG - RT - NoAWG	0,56	0,42	0,51	0,45	0,55	0,65	0,28	0,28	0,46
HoG - Multi SVMr - AWG	0,5	0,26	0,47	0,41	0,3	0,6	0,18	0,44	0,39
HoG - Multi SVMr - NoAWG	0,43	0,15	0,38	0,31	0,15	0,16	0,6	0,33	0,31
HoG - Single SVMC - No AWG	0,75	0,35	0,65	0,52	0,7	0,68	0,65	0,44	0,59
HoG - Single RT - No AWG	0,87	0,58	0,77	0,54	0,6	0,49	0,2	0,23	0,53
HoG - 4 ERT - AWG	0,9	0,39	0,78	0,31	0,81	0,19	0,26	0,37	0,5
HoG - 4 ERT - NoAWG	1	0	0,88	0	0,9	0	0,42	0	0,4
COV - ERT - AWG	0,3	0,15	0,7	0,29	0,2	0,21	0,26	0,28	0,3
COV - ERT - No AWG	0,33	0,1	0,53	0,29	0,3	0,18	0,28	0,28	0,28
COV - SVM - AWG	0,3	0,15	0,38	0,06	0,4	0,1	0,29	0,12	0,23
COV - SVM - NoAWG	0,2	0,1	0,32	0,11	0,45	0,06	0,33	0,15	0,21
COV2 - ERT - AWG	0,16	0,15	0,28	0,11	0,15	0,16	0,72	0,25	0,25
COV2 - ERT - NoAWG	0,16	0,15	0,28	0,11	0,15	0,16	0,72	0,25	0,25
COV2 - SVM - AWG	0,36	0,15	0,41	0,18	0,2	0,05	0,23	0,12	0,21
COV2 - SVM - NoAWG	0,33	0,05	0,31	0,18	0,15	0,05	0,26	0,12	0,18

Figure 6.18: Comparison of the proposed method with alternative solutions exploiting different classifiers and features

	E	NE	N	NW	W	SW	S	SE	Overall
HoG - ERT - AWG	0,95	0,57	0,76	0,52	0,86	0,55	0,64	0,36	0,65
Chen et al. [4]	0,65	0,37	0,71	0,53	0,7	0,59	0,41	0,36	0,55
Andriluka et al.- Max [3]	0,54	0,35	0,46	0,23	0,38	0,08	0,4	0,08	0,31
Andriluka et al. - SVM [3]	0,73	0,13	0,49	0,12	0,56	0,44	0,7	0,16	0,42
Andriluka et al.- SVM-adj [3]	0,71	0,22	0,29	0,18	0,85	0,18	0,5	0,29	0,35

Figure 6.19: Comparison of the proposed method with the state-of-the-art

to 50, the maximum depth for each tree was set to 20. For the SVMs, ν-SVM classifiers were used, with a standard RBF kernel. The parameters γ was set to 0.000407, ν set to 0.5. For the MoAWG, σ was set to 0.75. Lastly, Fig. 6.19 compares the proposed method against the one presented by Chen et al. [48], which uses a similar feature vector and a sparse representation technique, and against the methods presented by Andriluka et al. [11], which exploit banks of viewpoint specific part based detectors (linear SVMs) trained on the 8 orientation classes. In the first row, the results of the proposal are reported as reference. The second row shows the performance of [48], while the other rows contain three variants of the method by Andriluka et al. [11]. In the first case (referenced as Max in the table), the orientation is estimated as the maximum over the outputs of 8 specific detectors; in the second case (SVM), eight additional SVMs are trained on top of the viewpoint specific detectors using training images from one

Figure 6.20: Some sample of the orientation and position estimation steps on the 3DPeS dataset

class as positive samples and the remaining ones as negatives samples. Finally, in the last case (SVM-Adj) orientations are grouped into triplets of adjacent directions and 8 linear SVMs are trained on such groups. This outperforms all the previous solutions; in particular, the improvement with respect to the one presented by Chen *et al.* in [48] is around the 18%. Fig. 6.20 reports some qualitative results of the orientation estimation on the 3DPeS dataset.

Chapter 7

Conclusions

Between the plethoras of re-identification approaches presented in the scientific literature so far, only a handful of them try to exploit 3D information in the re-identification process. SARC3D, thorough this thesis, proves to be a very reliable, efficient and effective solution to this problem. Despite being a 3D method, hence apparently requiring lots of prior data for calibration and image registering, this thesis proves that 3D (or 2.5D) methods can be easily exploited even when full calibration and image registration is not entirely possible, through a clever design of the model itself and the algorithms exploited. Results both in real surveillance videos and in the proposed new benchmark datasets (3DpeS,ViSOR) are very promising for the future. When additional depth-data is also available, a new solution has been presented in this thesis, based on articulated 3D body models that spatially localize identifying patterns and colors on virtual bones. Experimental results on a dataset captured with the Microsoft Kinect prove the improvement obtained using articulated models instead of fixed containers (such as the SARC3D) or 2D body models. It is my belief, and indeed it has been proven here, that this new explored way based on 3D body models could be the starting point for future innovative solutions.

Appendix A

Multi-View People Surveillance Using 3D Information

In this appendix a novel tracking system is presented, that was the result of a collaboration with Ákos Utasi, Csaba Benedek, Tamás Szirányi
of the Computer and Automation Research Institute, Hungarian Academy of Sciences. It was originally presented in [197] and combined with SARC3D in [22].

A.1 People Localization and Tracking via 3D Marked Point Process model

The proposed method operates in a multi-camera system, and its inputs are the Tsai's calibration parameters [195] and the foreground masks extracted from each view using a Mixture of Gaussians (MoG) background model [191]. The key idea of this step is to simultaneously project the foreground pixels on the ground plane, and on a parallel plane shifted to the estimated height of the person, see Fig. A.1. If this estimation is correct, it can observe from a birds-eye viewpoint that the point of osculation of the silhouette's ground and head plane projections is the

ground position of the person. Since the heights of the people are unknown, the masks are projected on multiple planes having distances from the ground in the range of typical human sizes. Then the projections from multiple views are fused together, and searched for the optimal configuration in an iterative process using the above features and geometrical constraints.

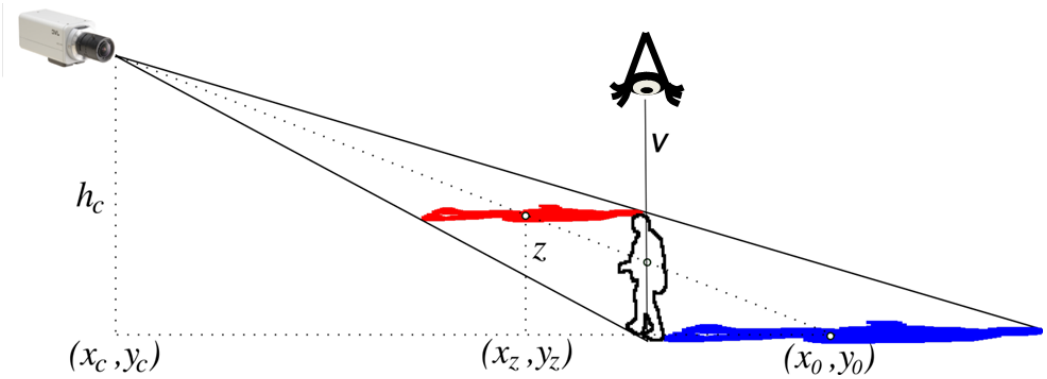


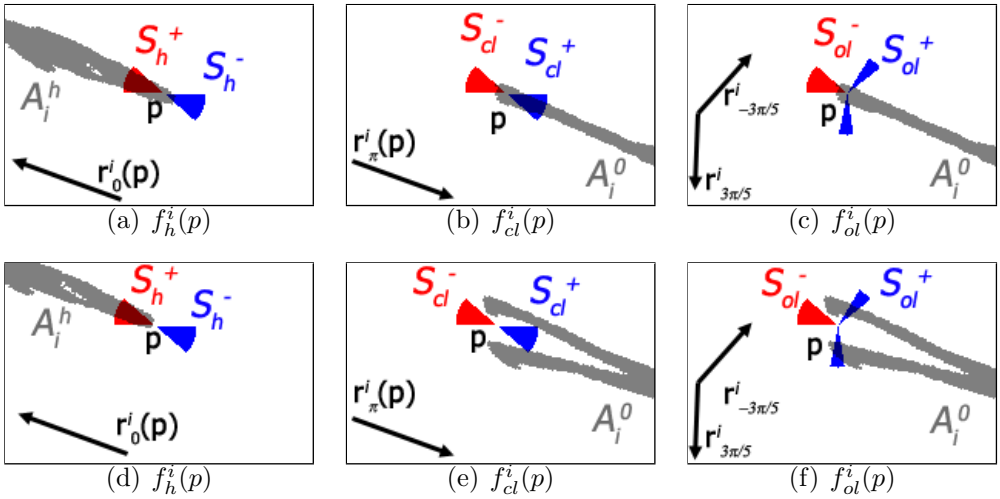
Figure A.1: The available camera calibration model is used for projecting the moving body silhouettes on the ground plane (blue) and on parallel planes (red) having different heights, source: [197].

A.1.1 Feature extraction

Our hypothesis on the location and height of a person is based on the 2D image formation of a 3D object in the conventional pinhole camera model. Lets consider in Fig. A.1 the person with height h, and project the silhouette on the P_0 ground plane (marked with blue) and on the P_z plane with the height of the person (*i.e.* $z = h$, marked with red). Also consider the v vertical axis of the person that is perpendicular to the P_0 plane. It can be observed that from this axis, the silhouette points projected to the $P_z|_{z=h}$ plane lie in the direction of the camera, while the silhouette print on P_0 is on the opposite side of v.

Based on the above observation a numerical feature is defined, which evaluates a given $[\mathbf{p}, h]$ object candidate. It is denoted by $\mathbf{r}_0^i(\mathbf{p})$ a unity vector, which points from $\mathbf{p}$ towards the ground position of the i-th camera on the P_0 plane, and by $\mathbf{r}_\varphi^i(\mathbf{p})$ the rotation of $\mathbf{r}_0^i(\mathbf{p})$ with angle φ. The foreground points of the i-th view

projected to the P_0 and P_h planes are also denoted by A_0^i (blue in Fig. A.1) and A_h^i (red), respectively.



(a) $f_h^i(p)$ (b) $f_{cl}^i(p)$ (c) $f_{ol}^i(p)$

(d) $f_h^i(p)$ (e) $f_{cl}^i(p)$ (f) $f_{ol}^i(p)$

Figure A.2: Calculation of the $f_h^i(p)$, $f_{cl}^i(p)$ and $f_{ol}^i(p)$ features in two selected positions, corresponding to a person with closed (top) and open (bottom) legs, respectively.

An object hypothesis $[\mathbf{p}, h]$ is relevant according to the i-th camera data if it jointly meets constraints about the *head* and *leg* positions. *On one hand*, we should find projected pixels on P_h (*i.e.* red prints) in the neighborhood of the $\mathbf{p}$ point in the $\mathbf{r}_0^i(\mathbf{p})$ direction, but penalize such silhouettes points in the opposite direction $\mathbf{r}_\pi^i(\mathbf{p})$. To measure this property, circular sectors S_h^+ and S_h^- around $\mathbf{p}$ directed into $\mathbf{r}_0^i(\mathbf{p})$ (red in Fig. A.2) and $\mathbf{r}_\pi^i(\mathbf{p})$ respectively are defined. The sectors have fixed arc and radius, which are parameters of the model. Then, following Fig. A.2(a) and (d), the *head* feature is computed as:

$$f_h^i(p) = \frac{\mathbf{Area}\big(A_h^i \cap S_h^+(p)\big) - \mathbf{Area}\big(A_h^i \cap S_h^-(p)\big)}{\mathbf{Area}\big(S_h^+(p)\big)} \ .$$

On the other hand, two different cases must be distinguished by the definition of the *leg* position constraint. People with closed *l*egs can be handled in an analogous manner to the *head* feature (see Fig. A.2(b)). Here S_{cl}^+ and S_{cl}^- sectors

correspond to $\mathbf{r}^i_\pi(\mathbf{p})$ and $\mathbf{r}^i_0(\mathbf{p})$ directions respectively, and

$$f^i_{cl}(p) = \frac{\mathbf{Area}\big(A^i_0 \cap S^+_{cl}(p)\big) - \mathbf{Area}\big(A^i_0 \cap S^-_{cl}(p)\big)}{\mathbf{Area}\big(S^+_{cl}(p)\big)} \ .$$

However, if the person is in the swing phase of the gait cycle the previous descriptor proves to be inaccurate (see Fig. A.2(e)). Instead, an *o*pen *l*eg feature was developed (see Fig. A.2(c) and A.2(f)), whose attractive region, S^+_{ol}, consists of two, half sized circular sectors corresponding to the directions $\mathbf{r}^i_{\pm 3\pi/5}(\mathbf{p})$. The repulsive sector, S^-_{ol} is constructed in the same way as S^-_{cl}. Then, $f^i_{ol}(p)$ feature term is derived similarly to $f^i_{cl}(p)$. Since it can be easily observed that for our purposes, the gait phase of each person can be fairly approximated either by the *c*losed or by the *o*pen leg states, the *joint leg* feature is obtained as $f^i_l(p) = \max(f^i_{cl}(p), f^i_{ol}(p))$. Finally, the *head* and *leg* features are truncated to take values in the $[0, \hat{f}]$ range, and are normalized by $\hat{f}$, which controls the ratio required to produce the maximal output.

If the object defined by the $[\mathbf{p}, h]$ parameters is completely visible for the i-th camera, both the $f^i_h(\mathbf{p})$ and $f^i_l(\mathbf{p})$ features should have *high* values. However, in the available views, some of the legs or heads may be partially or completely occluded by other pedestrians or static scene objects, which can strongly corrupt the feature values. Therefore a stronger feature can be constructed by averaging the responses of the N available cameras: $\bar{f}_h(\mathbf{p}) = 1/N \cdot \sum_{i=1}^N f^i_h(\mathbf{p})$, $\bar{f}_l(\mathbf{p}) = 1/N \cdot \sum_{i=1}^N f^i_l(p)$. Finally, the joint data feature $f(\mathbf{p}, h)$ is derived as $f(\mathbf{p}, h) = \sqrt{\bar{f}_h(\mathbf{p}) \cdot \bar{f}_l(\mathbf{p})}$.

A.1.2 3D Marked Point Process model

Since the goal of the proposed model is position and height estimation of the people, a person can be approximated by a cylinder u in the 3D scene, with a fixed radius R. The free parameters of the cylinder object are the center coordinate $\mathbf{p}$ on P_0 and the height h, *i.e.* $u = (\mathbf{p}, h)$ The aim of this methods is to to extract a configuration of n cylinder objects in the scene: $\omega = \{u_1, \ldots, u_n\}$ where n is also unknown.

The global input data ($\mathcal{D}$) of the model consists of the foreground masks and the calibration matrices. An input-dependent energy function on the configuration space is defined: $\Phi_{\mathcal{D}}(\omega)$, which assigns a *negative likelihood* value to each possible object population, and is divided into data dependent $J_{\mathcal{D}}$ and prior I parts:

$$\Phi_{\mathcal{D}}(\omega) = \sum_{u \in \omega} J_{\mathcal{D}}(u) + \gamma \cdot \sum_{\substack{u,v \in \omega \\ u \sim v}} I(u,v) \ , \tag{A.1}$$

where $J_{\mathcal{D}}(u) \in [-1,1]$, $I\,(u,v) \in [0,1]$ and γ is a weighting factor between the two terms. The $u \sim v$ relation holds if the two cylinders intersect.The optimal object population is derived as the maximum likelihood configuration estimate, *i.e.* $\omega_{\text{ML}} = \arg\min_{\omega \in \Omega} \left[\Phi_{\mathcal{D}}(\omega)\right]$.

In the next step, the I prior and $J_{\mathcal{D}}$ data-based potential functions should be defined appropriately so that the ω_{ML} configuration efficiently describes the group of people in the scene. First of all, configurations which contain many objects in the same or strongly overlapping positions need to be avoided. Therefore, the $I(u,v)$ *interaction* potentials realize a prior geometrical constraint: they penalize intersection between different object cylinders in the 3D model space:

$$I(u,v) = \mathbf{Area}\big(u \cap v\big) / \mathbf{Area}\big(u \cup v\big) \ . \tag{A.2}$$

On the other hand, the $J_{\mathcal{D}}(u)$ *unary* potential characterizes a proposed object candidate segment u depending on the image data, but independently of other objects. Cylinders with negative unary potentials are called *attractive objects*. Based on (A.1) the optimal population should consist of attractive objects exclusively: if $J_{\mathcal{D}}(u) > 0$, removing u from the configuration results in a lower $\Phi_{\mathcal{D}}(\omega)$ global energy.

At this point the $f_u = f(\mathbf{p}, h)$ feature is exploited in the Marked Point Process (MPP) model. Lets remember, that the f_u fitness function evaluates a person-hypothesis for u, so that 'high' f_u values correspond to efficient object candidates. For this reason, the feature domain is projected to $[-1,1]$ with a monotonously decreasing $Q(f_u, d_0)$ function: $J_{\mathcal{D}}(u) = Q(f_u, d_0) = 1 - f_u/d_0$, if $f_u < d_0$; $\exp\left(D^{-1} \cdot (f_u - d_0)\right) - 1$ otherwise. Here the d_0 parameter defines the

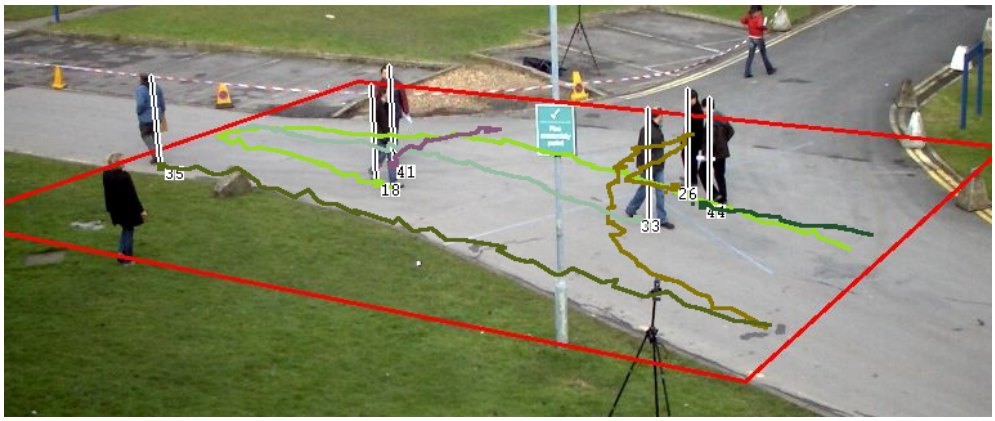
minimal value required for acceptance. Consequently, object u is attractive according to the $J_{\mathcal{D}}(u)$ term if and only if $f_u > d_0$, where the d_0 parameter defines the minimal value required for acceptance.

Since finding the optimal configuration according to (A.1) is NP-hard, quicker optimization techniques have to be used: the Multiple Birth and Death (MBD) algorithm [70] was choosen for this purpose, which evolves the population of people-cylinders by alternating randomized object generation (birth) and removal (death) steps in a simulated annealing framework, see details in [70, 197].

A.1.3 Short Term Tracking

labelsec:shortterm The output of the detection stage is the set of detections $\{u_n^t\}; n \in [1 \dots N^t]$ where N^t is the number of detected objects at time t. The short-term tracking system, instead, aims at creating and keeping updated a set of moving objects $\{o_p\}$. The current and future state of each object is estimated by means of a constant velocity Kalman filter. At each frame, a distance matrix between current detections and tracked objects is computed and, after a thresholding step, passed to a zero/one integer programming formulation for the assignments. The detection-to-object distance is computed using the Euclidean distance in the 3D space of the position and height of each object. The distance threshold has been set to a very low value in order to avoid wrong matches even if an over segmentation of the trajectories is introduced and handled by the long-term tracking system.

Unmatched detections are used to create new tracks only if they are localized in an entering area (to prune the wrong multiple detections which can be found in the center of the scene). Tracks without a matching detection, instead, are kept alive and updated using the Kalman prediction only. After a predefined time of inactivity or if their position exits from the scene the objects are definitively deleted. Fig.A.3(a) reports a qualitative example of the short-term tracking, with people id, position and trajectory superimposed. The red rectangle represents the region of interest (ROI). Broken trajectories and people entering again the scene after a while are managed by SARC3D.



(a)

(b)

Figure A.3: (a) Estimated positions and heights are represented by a line. The ids and trajectories are also superimposed using different color. The red area corresponds to the ROI. (b) The 3D body models are placed in the estimated ground positions, orientation is estimated from the trajectory.

Appendix B

Fast Background Initialization with Recursive Hadamard Transform

In this chapter, a new and fast technique for background estimation from cluttered image sequences is presented. Most of the background initialization approaches developed so far collect a number of initial frames and then require a slow estimation step which introduces a delay whenever it is applied. Conversely, the proposed technique redistributes the computational load among all the frames by means of a patch by patch preprocessing, which makes the overall algorithm more suitable for real-time applications. For each patch location a prototype set is created and maintained. The background is then iteratively estimated by choosing from each set the most appropriate candidate patch, which should verify a sort of frequency coherence with its neighbors. To this aim, the Hadamard transform has been adopted which requires less computation time than the commonly used DCT. Finally, a refinement step exploits spatial continuity constraints along the patch borders to prevent erroneous patch selections. The approach has been compared with the state of the art on videos from available datasets (ViSOR and CAVIAR), showing a speed up of about 10 times and an improved accuracy.

B.1 Introduction

Segmentation of foreground objects using motion information is a core aspect in many computer vision systems and in particular in automated visual surveillance. Commonly, a foreground/background pixel-wise classification algorithm is adopted for each frame, relying on a background model which should be kept updated correctly. A wide variety of algorithms for background modeling and updating have been proposed [168, 175]; among the others, Mixtures of Gaussian [190] or statistical models [59] have been widely adopted. However, background initialization is still a challenging problem, in particular when all the frames contain moving objects and the empty background is never seen as a whole. Background initialization should also be very effective in cluttered video with many moving objects and fast enough to be used in real time.

Often the problem of background initialization is neglected and it is directly merged into the estimation and update steps: starting from an unreliable model, errors are identified and corrected by analyzing the extracted foreground objects. For example, in [59] a rough classification of the foreground objects is provided with a motion-based validation step and the "ghosts" (i.e., regions of apparent motion but classified as foreground object due to a dirty background) are used to update the background. Object size, edges, optical flow or other features can be exploited to post process the detected foreground regions and to discard the erroneously detected objects [99, 162].

Broadly speaking, between three classes of background modeling algorithms can be distinguished, depending on the spatial level used. All the above described methods work at a region level, and usually are characterized by a high computational cost. Conversely, several methods try to solve the background initialization problem working at a pixel level, mainly exploiting the pixel intensity temporal constancy [137]. Even if these methods are very fast, they do not exploit spatial relations. To mitigate the problem, statistical models for each pixel [190, 192] or multiple background images have been proposed [32]. Finally, intermediate solutions work at a block (or patch) level [51, 176, 179].

Independently from the spatial level adopted, two approaches can be defined: recursive and non-recursive techniques. Recursive approaches maintain a single

background model that is updated with each new video frame. These techniques are generally computationally efficient and have minimal memory requirements. Non-recursive techniques, instead, maintain a buffer of previous video frames and estimate a background model based solely on the statistical properties of these frames. This causes non-recursive techniques to have higher memory requirements than recursive techniques. However, since they have explicit access to the most recent video frames they can model aspects of the data which can't be analyzed with recursive techniques.

In [51], for example, when enough frames have been collected, static blocks are selected as reference and directly included into the background; then the model is completed by iteratively adding blocks which satisfy spatial consistency and homogeneity constraints. Recently Reddy et al. [176] applied a similar approach, using the DCT coefficients as a core feature to check the homogeneity constraint. These methods are general and perform very well on different types of videos. However, they are computationally too expensive to be correctly applied in real time. For example, the method by Reddy et al. [176] contains an iterative block selection which prevents parallel solutions and introduces a long delay when it is executed.

In this appendix a new fast background initialization method is proposed, working at block level in a non-recursive way, specially conceived for achieving the best background model using the minimum number of frames as possible. Similarly to [176]each frame is split into blocks, producing a history of blocks and searching among them for the most reliable ones. In this last phase, the method works at a super-block level evaluating and comparing the frequency content of each block component. Differently from [176] which makes use of the DCT coefficients, the Hadamard Transform is adopted [100] which is faster and particularly suitable for this type of applications. In the next section a brief description of the Hadamard transform and it's properties is given. In section B.3 the proposed algorithm is described in detail, and results from real-life surveillance videos are reported in section B.4.

B.2 The Hadamard Transform

The Hadamard transform [100, 159] belongs to the generalized class of Fourier transforms and it is based on the homonym matrix. The Hadamard matrix is a square array whose elements can be ± 1 only and its rows (and columns) are mutually orthogonal. It's recursive definition is one of the more interesting properties for our purposes. The lowest-order Hadamard matrix $\mathbf{H}_1$ has size 1x1 and it is defined as $\mathbf{H}_1 = [1]$. The Hadamard matrices having order $N = 2^n$, with n integer, can be recursively defined. In particular, given the Hadamard matrix of order N, the Hadamard matrix of order $2N$ is defined as:

$$\mathbf{H}_{2N} = \begin{bmatrix} \mathbf{H}_N & \mathbf{H}_N \\ \mathbf{H}_N & -\mathbf{H}_N \end{bmatrix} \tag{B.1}$$

Fig. B.1 contains some example of Hadamard matrices of various orders.

A frequency interpretation of the Hadamard matrix can be given [171]. Along each row of the matrix, the frequency is related to the number of changes in sign. This frequency interpretation of the rows of a Hadamard matrix allows us to consider the rows to be equivalent to rectangular waves ranging between $\pm\mathbf{1}$ with a sub-period of $1/N$ units. Thus, in this context the Hadamard matrix merely performs the decomposition of a function by a set of rectangular waveforms rather than the cosine waveforms associated with the DCT.

The aforementioned structure of the Hadamard matrix shows a key feature that proves extremely useful: let $\mathbf{f}(x, y)$ be an image (or a patch extracted from it) of $2N \times 2N$ pixels. Its Hadamard transform, $\mathbf{F}(u, v)$, is given by the matrix product

$$\mathbf{F} = \mathbf{M} \cdot \mathbf{f} \cdot \mathbf{M} \tag{B.2}$$

where $\mathbf{M}$ is the Hadamard matrix of order $2N$.

Another way to compute $\mathbf{F}(u, v)$ is by means of the Hadamard transform of constitutive blocks. Let us decompose the image $\mathbf{f}(x, y)$ into four $N \times N$ blocks, called A, B, C, D respectively. The product of Eq. B.2 can be accordingly decomposed as reported in Eq. B.3:

$N=2$

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$N=4$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

$N=8$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

Figure B.1: Hadamard matrices of order $N = 2^n$

$$\mathbf{F} = \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & -\mathbf{H} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & -\mathbf{H} \end{bmatrix} \tag{B.3}$$

where $\mathbf{H}$ is the Hadamard matrix of order N. This leads to:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_{1,1} & \mathbf{f}_{1,2} \\ \mathbf{f}_{2,1} & \mathbf{f}_{2,2} \end{bmatrix} \tag{B.4}$$

where

$$\begin{aligned} \mathbf{f}_{1,1} &= \mathbf{HAH} + \mathbf{HBH} + \mathbf{HCH} + \mathbf{HDH} \\ \mathbf{f}_{1,2} &= \mathbf{HAH} - \mathbf{HBH} + \mathbf{HCH} - \mathbf{HDH} \\ \mathbf{f}_{2,1} &= \mathbf{HAH} + \mathbf{HBH} - \mathbf{HCH} - \mathbf{HDH} \\ \mathbf{f}_{2,2} &= \mathbf{HAH} - \mathbf{HBH} - \mathbf{HCH} + \mathbf{HDH}. \end{aligned} \tag{B.5}$$

The elements **HAH**,**HBH**,**HCH** and **HDH** in Eq. B.5 are the Hadamard transforms of the blocks A, B, C, D. Therefore, it's possible to compute the Hadamard transform of order **2N** from the four Hadamard transforms of order **N**, something that it's not possible in the case of the discrete cosine transform.

In addition to the recursive formulation, evaluating the Hadamard transform is faster than estimating the DCT. For example, let us consider a 8x8 block: one of the fastest DCT algorithms requires 94 multiplication and 454 additions [81], while the Hadamard transform only requires 384 integer additions for the same block (see [171] for details on the fast Hadamard computation algorithm).

Furthermore, a single 32x32 DCT (the usual size for the super-block in our algorithm) would require on average 18.560 real-arithmetic operation (multiplications and additions) [189]. By using the Hadamard transform that number is reduced to 10240 integer additions; finally by exploiting the Hadamard matrix structure by computing the 16x16 Hadamard transforms as soon as a frame is available, the number of operations required for the 32x32 transform is reduced to 3072 integer additions.

B.3 Recursive Hadamard Transform Background Initialization

Commonly to the other methods which estimate the background after collecting a statistically sufficient number of frames, the proposed technique analyze only the first T frames $I_t, t \in 1 \dots T$ of a video sequence, supposing that the number T of analyzed frames is set high enough to guarantee that each part of the background is visible in at least one frame.

Similarly to [176], each frame I_t is partitioned into disjoint square blocks of 16x16 pixels. Let $b_t^{i,j}$ be the block extracted from I_t at the position (i, j). From the above mentioned assumptions, it can be reasonably asserted that the final background image should be obtained by the composition of opportunely selected blocks.

To this aim, a three step algorithm is proposed. In the first step, each block lo-

B	C	D
A	X	E
H	G	F

Table B.1: 8-connected neighbors of block X

cation is independently analyzed and a set of representative blocks is constructed for each of them. The blocks belonging to sets which contain one element only are automatically selected and fixed as background. In other words, if a block is stable for all the first T frames, it is certainly a background block. Then, the second step aims at selecting the remaining blocks following a growing schema at a super-block level. New blocks are iteratively added to the background if they are similar enough to three neighbors which have been already included in the background.

The super-block should be constructed using the two 4-connected neighbors and the diagonal block between them. In table B.1, for example, block X can be estimated thanks to blocks A,B and C, or blocks C,D,E, and so on.

The block similarity is estimated evaluating the frequency coherence inside a super-block with the Recursive Hadamard Transform. Finally, all candidate blocks are checked again using a new refinement step to assure spatial continuity of the background image along the block borders. These three steps are described in the following subsections.

B.3.1 Block candidate sets

For each location (i, j), a representative set $\mathbf{R}^{i,j} = \left\{r_k^{i,j}\right\}$ of unique blocks should be extracted. Each element $r_k^{i,j}$ is associated to a weight $\mathbf{W}_k^{i,j}$, that denotes the number of occurrences in the image sequence. The blocks of the first frame are automatically inserted in their corresponding sets, with an initial weight set to 1. Then, for each frame $I_{2...T}$ and for each location, the corresponding block is compared with all the elements of the set, looking for the most similar item. If the new block $b_t^{i,j}$ is unique (i.e., it is different enough from all the other representatives) it is added to the set with weight 1; otherwise, it is used to update

the most similar block in the set through a weighted mean and the corresponding weight is incremented by 1.

As proposed in [176], the similarity between two blocks b_t and r_k is checked by means of the cross-correlation (Γ) and the MAD (Φ) coefficients, respectively computed as:

$$\Gamma(r_k, b_t) = \frac{\sum_{x=1}^{N}\sum_{y=1}^{N}[r(x,y) - \mu_{r_k}] \cdot [b(x,y) - \mu_{b_t}]}{\sigma_{r_k}\sigma_{b_t}} \tag{B.6}$$

$$\Phi(r_k, b_t) = \frac{1}{N^2}\sum_{x=1}^{N}\sum_{y=1}^{N}\|r_k(x,y) - b_t(x,y)\| \tag{B.7}$$

where μ and σ are the intra-block mean and standard deviation respectively. The two blocks $\mathbf{b}_t$ and $\mathbf{r}_k$ are considered similar if

$$\Gamma(r_k, b_t) > \alpha \quad \wedge \quad \Phi(r_k, b_t) < \beta \tag{B.8}$$

where α and β are empirically selected; β can also be automatically estimated using the following approach (as described in [176]). Using a short training video, the MAD coefficients between co-located blocks of successive frames are calculated. These values are sorted and only the central half of them are kept. Calling μ and σ their mean and standard deviation respectively, β can be set equal to $\mu + 2 \cdot \sigma$. This ensures that low MAD values (close or equal to zero) and high MAD values (arising due to object movements) are treated as outliers and thus ignored.

B.3.2 Frequency-based block selection

At time T, when all the frames have been analyzed, the background image BG can be effectively generated from the $\mathbf{R}^{i,j}$ sets. The background is firstly initialized using stable and unchanged blocks, which are characterized by $\mathbf{W}_k^{i,j} = T$. These blocks are also quickly identified from the sets with one element only. Instead,

for those sets with more than one members, the background block is chosen by analyzing every representative block $r_k^{i,j}$ and comparing them with their already chosen neighborhoods in the frequency domain.

The background block at the location (i, j) can be estimated if it is the only missing item in a 2x2 super-block. See fig. B.2 for an example: the X block need to be analyzed, given that the A,B,C blocks have been already estimated. A set of super-blocks can be constructed based on the elements of $\mathbf{R}^{i,j}$ and the known adjacent background blocks. The first super-block (called, base super-block) is constructed by forcing block X elements to zero, and filling blocks A,B,C with the known data. It's $2N$-Hadamard transform is computed from the N-Hadamard transforms of blocks A,B,and C (the Hadamard transform of a null block is a null matrix), resulting in a matrix $\mathbf{C}$ of size $\mathbf{M} \times \mathbf{M}$, with ($\mathbf{M} = \mathbf{2N}$). Then for each block in the representative set of X, a super-block is constructed by forcing A,B,C to zero and filling X with the block data. It's Hadamard transform is easily constructed, being the Hadamard transforms of blocks A,B and C a null matrices. This means that constructing the Hadamard transform of this super-block requires no arithmetic operations at all. The result is stored in a $\mathbf{M} \times \mathbf{M}$ matrix $\mathbf{D}_k$. A cost function is then defined as:

$$cost(k) = \left(\sum_{v=0}^{M-1} \sum_{u=0}^{M-1} |C(u, v) + D_k(u, v)| \right) \lambda_k \tag{B.9}$$

where $\lambda_k = e^{-\delta\omega_k}$, with $\delta \in [0, 1]$ and $\omega_k = \mathbf{W}_k / \sum_{k=1}^{S} \mathbf{W}_k$. The representative block which yields the minimum value of the cost function is assumed to be the best candidate as background block.

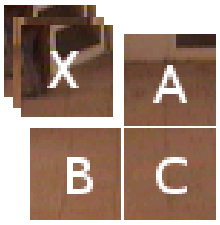


Figure B.2: A super-block of block X

B.3.3 Spatial continuity check and selection refinement

(a) Errors in the estimated background

(b) Corrected estimated background

Figure B.3: Estimated background before and after block selection correction

After all the previous steps, the chosen blocks are analyzed again to prevent some errors (as can be seen in Fig. B.3(a)). Actually, the frequency domain constraints embedded in Eq. B.9 are not enough to assure image continuity along each block border. In fact, blocks candidates can have a very similar frequency content but a different aspect, in particular this happens inside flat background parts.

The average gradient along each border is computed, and if the average gradient of two or more sides is greater than a given threshold γ, the selected block is discarded and a new background block is selected among the remaining unique representative blocks of the corresponding set, using the same algorithm described in section B.3.2. The threshold value γ has been empirically set using part of the CAVIAR dataset [46] (which yields the most errors in our experiments). Given the sets of correctly and erroneously selected blocks, γ is the mean value of the corresponding intra-set average gradients.

This final refinement allows to further improve the estimation accuracy.

Table B.2: Datasets Summary

Dataset	Subset	Num. of Videos	Description	Frames (Start - End)	Size	Parameters
CAVIAR	set 1	28	large indoor room	100 (300-400)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 10$, $\gamma = 60$
CAVIAR	set 2cor	26	indoor corridor side view	100 (0-100)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 10$, $\gamma = 60$
CAVIAR	set 2front	26	indoor corridor front view	100 (0-100)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 10$, $\gamma = 60$
ViSOR	Outdoor Unimore D.I.I. setup - Multicamera	28	outdoor large space	100 (100-200)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 4$, $\gamma = 60$
ViSOR	Indoor Domotic Unimore D.I.I. setup	16	indoor small room	100 (100-200)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 4$, $\gamma = 60$
ViSOR	Video for indoor people tracking with occlusions	6	indoor small room	100 (100-200)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 4$, $\gamma = 60$
ViSOR	Outdoor Unimore D.I.I. setup - Multicamera - disjoint views	14	outdoor large space	100 (200-300)	384*288	$\alpha = 0.8$, $\delta = 0.5$ $\beta = 4$, $\gamma = 60$

B.4 Experimental results

Experiments on a total of 144 surveillance videos were carried out, as reported in Table B.2. The starting frame of each sequence was chosen accordingly to avoid trivial conditions like uncluttered scenes and background-only frames. Few videos were also resized to 384×288 from their original dimensions to obtain comparable results.

RHT was compared with the DCT-based algorithm presented in [176] and a trivial median filter approach. The three methods have been implemented in C++, partially using the OpenCV libraries and the Imagelab processing libraries. Tests were performed on a 1.6 GHz dual core processor.

Parameters were set as follows: block size was set to **16x16** pixels, α was empirically set to **0.8** as well as δ set to **0.5**, β was automatically set to **10** for the CAVIAR dataset, and **4** for the ViSOR dataset, γ was set to **60**.

The DCT algorithm was able to elaborate images at 33 fps, but the actual background computation was done at the end of the process in 1506 ms (on average). This time is not compliant with a real time processing and the delay introduced after the frame T is considerable. The median background has the same drawback, with an even greater delay due to the sorting algorithm which should be performed on all the frames. Using the proposed approach, instead, the computational load is more balanced, images are processed at 27 fps, while the actual background computation at the end takes on average 97ms only. Table

Figure B.4: Example from two VISOR and two CAVIAR videos: (A,B) two random frames, (C) Estimated background using the median filter, (D) using the DCT based method of Reddy *et al.* ([176]), (E) Our proposed enhanced method

B.3 shows timing results for the three algorithms.

Table B.3: Timing results

	Frame Update (ms)	Background Estimation (ms)
Median	46	3133
DCT-based Method [176]	29	1506
RHT	**36**	**97**

To evaluate objectively the resulted background images, a methodology similar to the one presented in [99] was used, but extended to color images. In order to compare the results from the three algorithms the average error (AE) and the number of clustered error pixels (CEP) are used. AE is the average of the distances between the pixels of the estimated and true background, while CEP is the number of error pixels that are 4-connected to other error pixels. A pixel of the estimated background is considered an error pixel if the distance from the same pixel of the true background is greater than 20. Table B.4 shows the averaged

Table B.4: Averaged results using CAVIAR dataset

	Average error	Clustered error pixels
Median	16.00	5451
DCT-based Method [176]	14.12	3822
RHT	**12.55**	**2334**

Table B.5: Averaged results using ViSOR dataset

	Average error	Clustered error pixels
Median	11.080	1929
DCT-based Method [176]	13.55	1807
RHT	**12.62**	**968**

results for the CAVIAR dataset [46], while table B.5 shows results for the ViSOR dataset [202].

Fig. B.4 shows example results from four video sequences, the first two from the ViSOR dataset and the last two from the CAVIAR dataset.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. *Comput. Vis. Image Underst.*, 73(3):428–440, 1999. 8

[2] Y. Agiomyrgiannakis and Y. Stylianou. Wrapped gaussian mixture models for modeling and high-rate quantization of phase data of speech. *T-ASLP*, 17(4):775 –786, may 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2008229. 72

[3] Alexandre Alahi, Pierre Vandergheynst, Michel Bierlaire, and Murat Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, 2010. ISSN 1077-3142. 13, 20, 24

[4] 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, sept. 2012. ISSN 1751-9632. doi: 10.1049/iet-cvi.2011.0140. 13, 34

[5] A.B. Albu, D. Laurendeau, S. Comtois, D. Ouellet, P. Hebert, A. Zaccarin, M. Parizeau, R. Bergevin, X. Maldague, R. Drouin, S. Drouin, N. Martel-Brisson, F. Jean, H. Torresan, L. Gagnon, and F. Laliberte. MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras. In *Proc. of ICPR*, pages 924–928. IEEE, 2006. ISBN 0-7695-2521-0. vii, 11, 15, 21, 29, 49

[6] Saad Ali, Omar Javed, Neils Haering, and Takeo Kanade. Interactive retrieval of targets for wide area surveillance. In *Proc. of the ACM Interna-*

tional Conference on Multimedia, MM '10, pages 895–898, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. 17, 24, 33, 85

[7] Tauseef Ali, Raymond Veldhuis, and Luuk Spreeuwers. Forensic face recognition: A survey, 2010. URL http://doc.utwente.nl/75541/. 45, 46

[8] Enrique Amig, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12:461–486, 2009. ISSN 1386-4564. 43

[9] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, pages 1 –8, june 2008. 6, 31

[10] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. of CVPR*, pages 1014 –1021, june 2009. doi: 10.1109/CVPR.2009.5206754. 29

[11] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *Proc. of CVPR*, pages 623 –630, June 2010. doi: 10.1109/CVPR.2010.5540156. 30, 48, 69, 99, 101, 104

[12] Nadeem Anjum and Andrea Cavallaro. Trajectory Association and Fusion across Partially Overlapping Cameras. In *Proc. of AVSS*, pages 201–206, 2009. 13, 20

[13] Sanjeev Arulampalam, Simon Maskell, and Neil Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002. 62

[14] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning implicit transfer for person re-identification. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *ECCV Workshops (1)*, volume 7583 of *Lecture Notes in Computer Science*, pages 381–390. Springer, 2012. ISBN 978-3-642-33862-5. 78

[15] Kheir-Eddine Aziz, Djamel Merad, and Bernard Fertil. People re-identification across multiple non-overlapping cameras system by appear-

ance classification and silhouette part segmentation. In *Proc. of AVSS*, pages 303 –308, September 2011. vii, 16, 23

[16] B. Babenko, Ming-Hsuan Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Proc. of CVPR*, pages 983 –990, june 2009. 26, 27

[17] C. Bahlmann. Directional features in online handwriting recognition. *Pattern Recognition*, 39(1):115–125, January 2006. 73

[18] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Proc. of AVSS*, pages 179 –184, 30 2011-sept. 2 2011. 16, 36

[19] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. of AVSS*, pages 435–440, 2010. vii, 11, 16, 21, 23, 27, 30, 32, 49

[20] D.J. Balding. *Weight-of-Evidence for Forensic DNA Profiles*. Wiley, 2005. 45

[21] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3d body model construction and matching for real time people re-identification. In *Proc. of Eurographics Italian Chapter Conference 2010 (EG-IT 2010)*, Genova, Italy, November 2010. 22, 27, 47

[22] Davide Baltieri, Akos Utasi, Roberto Vezzani, Benedek Csaba, Tamas Sziranyi, and Rita Cucchiara. Multi-view people surveillance using 3d information. In *Proceedings of the Eleventh International Workshop on Visual Surveillance 2011*, pages 1817–1824, Barcelona, Spain, November 2011. 60, 107

[23] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *Proc. of IEEE Int. Conf. on Image Analysis and Processing*, pages 197–206, Ravenna, Italy, September 2011. vii, 16, 39, 40, 47, 69, 85, 86, 96, 99

[24] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proc. of the 1st Int. ACM Workshop on Multimedia Access to 3D Human Objects*, Scottsdale, Arizona, USA, November 2011. vii, 39, 40, 41, 69, 85, 99

[25] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. People orientation recognition by mixtures of wrapped distributions on random trees. In *Proceedings of the 12th European Conference on Computer Vision*, Firenze, Italy, October 2012. 67, 99

[26] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *ECCV Workshops (1)*, volume 7583 of *Lecture Notes in Computer Science*, pages 433–442. Springer, 2012. ISBN 978-3-642-33862-5. 24, 34, 77

[27] Martin Bauml and Rainer Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *Proc. of AVSS*, pages 291–296, 30 2011-sept. 2 2011. 16

[28] L. Bazzani, M. Farenzena, A. Perina, V. Murino, and M. Cristani. Multiple-shot person re-identification by hpe signature. In *Proc. of ICPR*, pages 1413–1416, 2010. 36, 39

[29] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 2012. ISSN 0167-8655. doi: 10.1016/j.patrec.2011.11.016. URL http://www.sciencedirect.com/science/article/pii/S0167865511004065. 15, 21, 27, 86

[30] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *Proc. of CVPR*, volume 1, pages 744 – 750, june 2006. 12

[31] Luca Bertelli, Tianli Yu, Diem Vu, and Burak Gokturk. Kernelized structural svm learning for supervised object segmentation. In *Proceedings of*

IEEE Conference on Computer Vision and Pattern Recognition 2011, 2011. 48, 50

[32] A. Bevilacqua, L. Di Stefano, and A. Lanza. An effective multi-stage background generation algorithm. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance,*, pages 388–393, Los Alamitos, CA, USA, 2005. IEEE Computer Society. ISBN 0-7803-9385-6. doi: http://doi.ieeecomputersociety.org/10.1109/AVSS.2005.1577300. 115

[33] S.T. Birchfield and Sriram Rangarajan. Spatiograms versus histograms for region-based tracking. In *Proc. of CVPR*, volume 2, pages 1158 – 1163 vol. 2, June 2005. doi: 10.1109/CVPR.2005.330. 24

[34] N.D. Bird, O. Masoud, N.P. Papanikolopoulos, and A. Isaacs. Detection of Loitering Individuals in Public Transportation Areas. *IEEE Transactions. on Intelligent Transportation Systems*, 6(2):167–177, June 2005. ISSN 1524-9050. vii, 11, 17, 19, 29, 31, 34, 49

[35] J. Black, T. Ellis, and P. Rosin. Multi view image surveillance and tracking. In *Proc. of Workshop on Motion and Video Computing, 2002*, pages 169–174. IEEE Comput. Soc, 2002. ISBN 0-7695-1860-5. 15, 20

[36] J. Black, T.J. Ellis, and D. Makris. Wide area surveillance with a multi camera network. *IEE Seminar Digests*, 2004(10426):21–25, 2004. doi: 10.1049/ic:20040092. URL http://link.aip.org/link/abstract/IEESEM/v2004/i10426/p21/s1. 25

[37] R. Bowden and P. KaewTraKulPong. Towards automated wide area visual surveillance: tracking objects between spatially-separated, uncalibrated views. *IEE Proceedings on Vision, Image and Signal Processing*, 152(2): 213 – 223, april 2005. ISSN 1350-245X. doi: 10.1049/ip-vis:20041233. 25

[38] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Proc. of ICCV*, pages 1515 –1522, 29 2009-oct. 2 2009. 31

[39] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *Proc. of CVPR*, pages 1273 –1280, june 2011. 12, 31

[40] B. Brosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. of BMVC*, pages 21.1–11, 2010. 16, 23, 33, 36, 85

[41] Q. Cai and J.K. Aggarwal. Tracking human motion using multiple cameras. In *Proc. of ICPR*, volume 3, pages 68 – 72, 1996. 12, 19

[42] Q. Cai and J.K. Aggarwal. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In *Proc. of ICCV*, pages 356–362. Narosa Publishing House, 1998. ISBN 81-7319-221-9. 18, 19, 20

[43] Q. Cai and J.K. Aggarwal. Tracking human motion in structured environments using a distributed-camera system. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 21(11):1241–1247, 1999. ISSN 01628828. 15, 19, 22

[44] S. Calderara, R. Cucchiara, and A. Prati. Bayesian-competitive consistent labeling for people surveillance. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 30(2):354–360, February 2008. 14, 18, 19, 20, 23, 30, 59

[45] Simone Calderara, Andrea Prati, and Rita Cucchiara. Mixtures of von mises distributions for people trajectory shape analysis. *IEEE Transactions on Circuits Syst. Video Technol.*, 21(4):457–471, April 2011. 73

[46] CAVIAR. Ec funded caviar project/ist 2001 37540. Website. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/. 123, 126

[47] T.-H. Chang and S. Gong. Tracking multiple people with a multi-camera system. In *Proc. IEEE Workshop Multi-Object Tracking*, pages 19–26. IEEE Comput. Soc, 2001. ISBN 0-7695-1171-6. 15

[48] Cheng Chen, A. Heili, and J. Odobez. Combined estimation of location and body pose in surveillance video. In *Proc. of AVSS*, pages 5 –10, 30 2011-sept. 2 2011. doi: 10.1109/AVSS.2011.6027284. 101, 104, 105

[49] Kuan-Wen Chen, Chih-Chuan Lai, Yi-Ping Hung, and Chu-Song Chen. An adaptive learning method for target tracking across multiple cameras. In *Proc. of CVPR*, pages 1 –8, june 2008. 14

[50] 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. 22, 29, 86

[51] Andrea Colombari, Andrea Fusiello, and Vittorio Murino. Background initialization in cluttered sequences. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 197, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2646-2. doi: http://dx.doi.org/10.1109/CVPRW.2006.40. 115, 116

[52] Alberto Colombo, James Orwell, and Sergio Velastin. Colour Constancy Techniques for Re-Recognition of Pedestrians from Multiple Surveillance Cameras. In *Proc. of Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008*, Marseille, France, 2008. 14, 23, 26

[53] C. Colombo, A. Del Bimbo, and A. Valli. A real-time full body tracking and humanoid animation system. *Parallel Comput.*, 34:718–726, December 2008. ISSN 0167-8191. doi: 10.1016/j.parco.2008.09.004. URL `http://portal.acm.org/citation.cfm?id=1464503.1464550`. 30, 48

[54] D. N. Truong Cong, L. Khoudour, and C. Achard. People reacquisition across multiple cameras with disjoint views. In *Proc. of Int. Conf. on Image and Signal Processing*, ICISP'10, pages 488–495, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-13680-X, 978-3-642-13680-1. 20

[55] D. N. Truong Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray. People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374, 2010. ISSN 0165-1684. 13, 20, 21

[56] D. Conte, P. Foggia, G. Percannella, and M. Vento. A multiview appearance model for people re-identification. In *Proc. of AVSS*, pages 297 –302, 30 2011-sept. 2 2011. 13

[57] Dalia Coppi, Simone Calderara, and Rita Cucchiara. Appearance tracking by transduction in surveillance scenarios. In *Proc. of AVSS*, September 2011. 31

[58] R. Cucchiara, C. Grana, G. Neri, M. Piccardi, and A. Prati. The sakbot system for moving object detection and tracking. In *Video-based Surveillance Systems - Computer Vision and Distributed Processing*. Kluwer Academic, 2001. 48

[59] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, October 2003. 60, 115

[60] R. Cucchiara, A. Prati, and R. Vezzani. Object segmentation in videos from moving camera with mrfs on color and motion features. In *Proc. of CVPR*, volume 1, pages 405–410, 2003. 48

[61] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. of ECCV*. Springer, 2006. 2

[62] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, volume 1, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2372-2. doi: http://dx.doi.org/10.1109/CVPR.2005.177. 68, 69, 71, 72

[63] A. Dantcheva, J.-L. Dugelay, and P. Elia. Soft biometrics systems: Reliability and asymptotic bounds. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1 –6, sept. 2010. doi: 10.1109/BTAS.2010.5634534. 24

[64] Antitza Dantcheva and Jean-Luc Dugelay. Frontal-to-side face re-identification based on hair, skin and clothes patches. In *Proc. of AVSS*, pages 309 –313, 30 2011-sept. 2 2011. 16, 24

[65] Antitza Dantcheva, Carmelo Velardo, Angela D'Angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification - new trends and challenges. *Multimedia Tools and Applications*, 51(2):739–777, 2011. 13, 24

[66] Icaro Oliveira de Oliveira and Jose Luiz Souza Pio. People Reidentification in a Camera Network. In *Proc. of 2nd Int. Conf. on Computer Science and its Applications*, pages 1–8. IEEE, December 2009. ISBN 978-1-4244-4945-3. 17, 24

[67] K. Delac and M. Grgic. A survey of biometric recognition methods. In *Proc. of Int. Symposium Electronics in Marine, ELMAR*, pages 184 – 193, june 2004. 24

[68] Giovanni Denina, Bir Bhanu, Hoang Thanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda. *VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication*, pages 335–347. Springer London, 2011. ISBN 978-0-85729-127-1. 38, 39

[69] Simon Denman, Clinton Fookes, Alina Bialkowski, and Sridha Sridharan. Soft-biometrics: Unconstrained authentication in a surveillance environment. In *Proc. of the 2009 Digital Image Computing: Techniques and Applications*, DICTA '09, pages 196–203, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3866-2. 18, 24

[70] X. Descombes, R. Minlos, and E. Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *J. of Math. Imaging and Vision*, 33(3):347–359, 2009. 112

[71] Jonathan Deutscher and Ian Reid. Articulated body motion capture by stochastic search. *Int. J. Comput. Vision*, 61(2):185–205, February 2005. ISSN 0920-5691. doi: 10.1023/B:VISI.0000043757.18370.9c. URL http://dx.doi.org/10.1023/B:VISI.0000043757.18370.9c. 33

[72] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Proceedings of the 10th Asian conference on Computer vision - Volume Part IV*, ACCV'10, pages 501–512, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-19281-4. URL http://dl.acm.org/citation.cfm?id=1966111.1966152. 27

[73] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. of CVPR*, pages 304–311, June 2009. 69

[74] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P.L. Mazzeo. A semi-automatic system for ground truth generation of soccer video sequences. In *Proc. of AVSS*, pages 559 –564, sept. 2009. doi: 10.1109/AVSS.2009.69. 38, 39

[75] Gianfranco Doretto, Thomas Sebastian, Peter H. Tu, and Jens Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *J. Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011. vii, 22, 23, 33

[76] Arnaud Doucet. On sequential simulation-based methods for bayesian filtering. Technical report, 1998. 62

[77] Helin Dutagaci, Blent Sankur, and Erdem Yrk. Comparative analysis of global hand appearance-based person recognition. *J. Electronic Imaging*, 17(1):1–19, 2008. 24

[78] T.J. Ellis and J. Black. A multi-view surveillance system. In *Proc. of IEE Symposium on Intelligence Distributed Surveillance Systems,*, pages 11/1 – 11/5, 2003. 20

[79] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *Proc. of ICCV*, October 2007. 37

[80] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of CVPR*, pages 2360–2367, June 2010. ISBN 978-1-4244-6984-0. vii, 11, 16, 21, 24, 27, 29, 31, 33, 36, 49, 85, 86, 89, 95, 96

[81] E. Feig and S. Winograd. Fast algorithms for the discrete cosine transform. *IEEE Transactions on Signal Processing*, 40(9):2174 –2193, sep 1992. ISSN 1053-587X. doi: 10.1109/78.157218. 119

[82] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. Wiley, 1968. ISBN 0471257087. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0471257087. 65

[83] Mika Fischer, Hazim Ekenel, and Rainer Stiefelhagen. Person re-identification in tv series using robust face recognition and user feedback. *Multimedia Tools and Applications*, 55:83–104, 2011. ISSN 1380-7501. 10.1007/s11042-010-0603-2. 16, 19, 24, 31

[84] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 1 edition, August 2002. ISBN 0130851981. 5

[85] Frontex. Application of surveillance tools to border surveillance - concept of operations. online, 2011. URL http://ec.europa.eu. 3, 43, 45

[86] Tarak Gandhi and Mohan Trivedi. Panoramic Appearance Map (PAM) for Multi-camera Based Person Re-identification. In *Proc. of AVSS*, pages 78–78. IEEE, November 2006. ISBN 0-7695-2688-8. vii, 14, 22, 23, 27, 30, 33

[87] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. 68, 72

[88] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person Reidentification Using Spatiotemporal Appearance. In *Proc. of CVPR*, volume 2, pages 1528–1535, 2006. ISBN 0769525970. 17, 20

[89] A. Gijsenij, T. Gevers, and J. van de Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475 –2489, sept. 2011. ISSN 1057-7149. doi: 10.1109/TIP.2011.2118224. 25

[90] Andrew Gilbert and Richard Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *Proc. of ECCV*, pages 125–136, 2006. 12, 25, 26

[91] Haifeng Gong, J. Sim, M. Likhachev, and Jianbo Shi. Multi-hypothesis motion planning for visual object tracking. In *Proc. of ICCV*, pages 619 –626, nov. 2011. 19, 29

[92] J. Gonzalez-rodriguez, J. Fierrez-aguilar, and J. Ortega-Garcia. Forensic identification reporting using automatic speaker recognition systems. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, pages 93–96, 2003. 45

[93] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, Dec 2007. 2

[94] H. Grabner, J. Matas, L. Van Gool, and P. Cattin. Tracking the invisible: Learning where the object might be. In *Proc. of CVPR*, pages 1285 –1292, june 2010. 27

[95] Douglas Gray and Hai Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Proc. of ECCV*, pages 262–275, Berlin, Heidelberg, 2008. Springer-Verlag. 14, 23, 24, 36, 85, 89, 90

[96] Douglas Gray, S Brennan, and H Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *Proc. of 10th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2007. vii, 36, 37, 38, 45, 84

[97] Giovanni Gualdi, Andrea Prati, and Rita Cucchiara. A multi-stage pedestrian detection using monolithic classifiers. In *Proc. of AVSS*, September 2011. 2

[98] Giovanni Gualdi, Andrea Prati, and Rita Cucchiara. Multi-stage particle windows for fast and accurate object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 2011. 48, 50

[99] D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons, and A.K. Jain. A background model initialization algorithm for video surveillance. In *Proc. of ICCV*, volume 1, pages 733–740, 2001. 115, 125

[100] J. Hadamard. Resolution d'une question relative aux determinants. *Bull. Sci. Math.*, 17:240–246, 1893. 116, 117

[101] Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu, and Bruno Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proc. of Int. Conf. on Distributed Smart Cameras*, pages 1–6. IEEE, September 2008. ISBN 978-1-4244-2664-5. 17, 22, 23, 32, 44

[102] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 19

[103] L. Havasi, Z. Szlavik, and T. Sziranyi. Eigenwalks: walk detection and biometrics from symmetry patterns. In *Proc. of ICIP*, pages III–289. IEEE, 2005. ISBN 0-7803-9134-9. 24

[104] Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 780–793. Springer, 2012. ISBN 978-3-642-33782-6. 16, 23, 25, 28, 78, 81

[105] Lei Hu, Shuqiang Jiang, Qingming Huang, and Wen Gao. People re-detection using Adaboost with sift and color correlogram. In *Proc. of ICIP*, pages 1348–1351. IEEE, 2008. 14, 23

[106] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 28(4): 663–671, 2006. 14, 19, 20, 23

[107] Timothy Huang and Stuart Russell. Object Identification: A Bayesian Analysis with Application to Traffic Surveillance. *Artificial Intelligence*, 103:1–17, 1998. 23

[108] Y. Hyodo, S. Yuasa, K. Fujimura, T. Naito, and S. Kamijo. Pedestrian tracking through camera network for wide area surveillance. In *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics*, pages 656–661. IEEE, October 2008. ISBN 978-1-4244-2383-5. 14, 20

[109] Anil K. Jain, Sarat C. Dass, Karthik Nandakumar, and Karthik N. Soft biometric traits for personal recognition systems. In *Proceedings of International Conference on Biometric Authentication, Hong Kong*, pages 731–738, 2004. 5, 24

[110] O. Javed and K. Shafique. Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras. In *Proc. of CVPR*, pages 26–33. IEEE, 2005. ISBN 0-7695-2372-2. 26, 32

[111] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008. ISSN 1077-3142. 14, 20, 25

[112] Chang Jing-Ying, Wang Tzu-Heng, Chien Shao-Yi, and Chen Liang-Gee. Spatial-temporal consistent labeling for multi-camera multi-object surveillance systems. In *Proc. of IEEE Int. Symposium on Circuits and Systems*, pages 3530–3533. IEEE, May 2008. ISBN 978-1-4244-1683-7. 14

[113] Nebojsa Jojic, Brendan J. Frey, and Anitha Kannan. Epitomic analysis of appearance and shape. In *Proc. of ICCV*, pages 34–43, 2003. 27

[114] Kai Jungling and M. Arens. View-invariant person re-identification with an implicit shape model. In *Proc. of AVSS*, pages 197 –202, 30 2011-sept. 2 2011. 13

[115] Kai Jungling and Michael Arens. Local Feature Based Person Reidentification in Infrared Image Sequences. In *Proc. of AVSS*, pages 448–455, 2010. 13, 19, 22, 45

[116] Jinman Kang, Isaac Cohen, and Gerard Medioni. Persistent Objects Tracking Across Multiple Non Overlapping Cameras. In *IEEE Workshop on Motion and Video Computing (WACV/MOTION'05)*, volume 2, pages 112–119. IEEE, January 2005. ISBN 0-7695-2271-8. 24

[117] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Proc. of CVPR*, pages 253–259. IEEE Comput. Soc, 1999. ISBN 0-7695-0149-4. 15, 20

[118] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360, October 2003. ISSN 0162-8828. 15, 18, 19, 20, 23, 30

[119] Saad M Khan and Mubarak Shah. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 31(3):505–19, March 2009. ISSN 0162-8828. 19, 20, 22

[120] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for EasyLiving. In *Proc. Third IEEE Int. Workshop on Visual Surveillance*, pages 3–10. IEEE Comput. Soc, 2000. ISBN 0-7695-0698-4. 15, 22

[121] Cheng-Hao Kuo and Ram Nevatia. How does person identity recognition help multi-person tracking. In *Proc. of CVPR*, pages 1217–1224, 2011. 12

[122] Cheng-Hao Kuo, Chang Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *Proc. of CVPR*, pages 685 –692, june 2010. 12, 26, 28

[123] Michel Lantagne, Marc Parizeau, and Robert Bergevin. VIP: Vision tool for comparing Images of People. In *Vision Interface*, 2003. vii, 11, 18, 21, 29

[124] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *First International ECCV Workshop on Re-Identification (ReID 2012)*, volume 7583 of *Lecture Notes in Computer Science*, pages 402–412. Springer, 2012. ISBN 978-3-642-33862-5. 15, 24, 34

[125] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 22:758–767, 2000. 20

[126] Sung Chun Lee and Ram Nevatia. Robust camera calibration tool for video surveillance camera in urban environment. In *Computer Vision and Pattern Recognition*, pages 62–67, 2011. doi: 10.1109/CVPRW.2011.5981777. 49

[127] Valerie Leung, James Orwell, and Sergio A. Velastin. Performance evaluation of re-acquisition methods for public transport surveillance. pages 705–712. IEEE, December 2008. ISBN 978-1-4244-2286-9. 44

[128] Anping Li, Zhongliang Jing, and Shiqiang Hu. Robust observation model for visual tracking in particle filter. *AEU - International Journal of Electronics and Communications*, 61(3): 186 – 194, 2007. ISSN 1434-8411. doi: DOI:10.1016/j.aeue.2006.03.009. URL http://www.sciencedirect.com/science/article/B7GWW-4JXY7FF-1/2/be7c74d6280805709d58e2b84d67f4e9. 64

[129] Qinghu Li, Qiming Chen, Tao Yu, and Wei Liu. A P2P Camera System with New Consistent Labeling Method Involving Only Simple Geometric Operations. In *Proc. of 11th IEEE Int. Symposium on Multimedia*, pages 52–56. IEEE, December 2009. ISBN 978-1-4244-5231-6. 13

[130] Wei Li, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Common-near-neighbor analysis for person re-identification. In *Internationl Conference on Image Processing*, pages 1621–1624, October 2012. 33

[131] Yuan Li, Chang Huang, and R. Nevatia. Learning to associate: Hybrid-boosted multi-target tracker for crowded scene. In *Proc. of CVPR*, pages 2953 –2960, june 2009. 28

[132] Guoyun Lian, Jianhuang Lai, and Yang Gao. People consistent labeling between uncalibrated cameras without planar ground assumption. In *Proc. of ICIP*, pages 733–736. IEEE, September 2010. ISBN 978-1-4244-7992-4. 13, 30

[133] Zhe Lin and Larry S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *Proc. of 4th Int. Symposium on Advances in Visual Computing*, pages 23–34, 2008. 17

[134] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *First International ECCV Workshop on Re-Identification (ReID 2012)*, volume 7583 of *Lecture Notes in Computer Science*, pages 391–401. Springer, 2012. ISBN 978-3-642-33862-5. 15, 33

[135] Kun Liu and Jie Yang. Recognition of People Reoccurrences Using Bag-Of-Features Representation and Support Vector Machine. In *Proc. of Chinese Conf. on Pattern Recognition*, pages 1–5. IEEE, November 2009. ISBN 978-1-4244-4199-0. 17, 22, 29, 86

[136] Y R Loke, Pankaj Kumar, Surendra Ranganath, and W M Huang. Object Matching Across Multiple Non-overlapping Fields of View Using Fuzzy Logic. *Acta Automatica Sinica*, 36(6):978–987, 2006. 14

[137] W. Long and Yee-Hong Yang. Stationary background generation: an alternative to the difference of two images. *Pattern Recogn.*, 23(12):1351–1359, 1990. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/0031-3203(90)90081-U. 115

[138] Fengjun Lv, Tao Zhao, and Ramakant Nevatia. Camera calibration from video of a walking human. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1513–1518, 2006. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.178. 49

[139] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *ECCV Workshops (1)*, volume 7583 of *Lecture Notes in Computer Science*, pages 413–422. Springer, 2012. ISBN 978-3-642-33862-5. 78

[140] Christopher Madden, Eric Dahai Cheng, and Massimo Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18(3):233, 2007. ISSN 0932-8092. 14, 23, 32

[141] F. Madrigal and J.-B. Hayet. Multiple view, multiple target tracking with principal axis-based data association. In *Proc. of AVSS*, pages 185 –190, 30 2011-sept. 2 2011. 13, 30

[142] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Proc. of CVPR*, pages 205–210. IEEE, 2004. ISBN 0-7695-2158-4. 20

[143] K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3):349–393, 1975. doi: 10.2307/2984782. 72

[144] Riccardo Mazzon, Syed Fahad Tahir, and Andrea Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, 33(14): 1828 – 1837, 2012. ISSN 0167-8655. doi: 10.1016/j.patrec.2012.02.014. URL http://www.sciencedirect.com/science/article/pii/S0167865512000554. ¡ce:title¿Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context¡/ce:title¿. 13, 20, 33

[145] Xue Mei and Haibin Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 33(11):2259 –2272, nov. 2011. ISSN 0162-8828. 26

[146] Michael Metternich, Marcel Worring, and Arnold Smeulders. Color Based Tracing in Real-Life Surveillance Data. *Trans. on Data Hiding and Multimedia Security V*, 6010:18–33, 2010. 16, 24, 26, 31, 85

[147] D. Meuwly. Forensic individualization from biometric data. *Science and Justice*, 46(4):205 – 213, 2006. ISSN 1355-0306. 45

[148] Florica Mindru, Tinne Tuytelaars, Luc Van Gool, and Theo Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94(1):3, 2004. ISSN 1077-3142. 23

[149] K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi, and H. Nakajima. An effective approach for iris recognition using phase-based image matching. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 30(10):1741 –1756, oct. 2008. ISSN 0162-8828. 24

[150] Eduardo Monari, Jochen Maerker, and Kristian Kroschel. A Robust and Efficient Approach for Human Tracking in Multi-camera Systems. In *Proc. of AVSS*, pages 134–139. IEEE, September 2009. ISBN 978-1-4244-4755-8. 17, 20, 29

[151] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, September 2003. ISSN 00313203. 18, 23

[152] A.T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *Proc. of AVSS*, pages 476 –481, September 2007. doi: 10.1109/AVSS.2007.4425357. 44

[153] A. Nilski. Evaluating multiple camera tracking systems - the i-lids 5th scenario. In *Security Technology, 2008. ICCST 2008. 42nd Annual IEEE International Carnahan Conference on*, pages 277 –279, oct. 2008. doi: 10.1109/CCST.2008.4751314. 36, 38

[154] Chaowei Niu and Eric Grimson. Recovering non-overlapping network topology using far-field vehicle tracking data. *Proc. of ICPR*, 4:944–949, 2006. ISSN 1051-4651. 18

[155] J.M. Odobez and P. Bouthemy. Detection of multiple moving objects using multiscale mrf with camera motion compensation. In *Proc. of ICIP*, volume 2, pages 257–261, 1994. 33

[156] Omar Oreifej, Ramin Mehran, and Mubarak Shah. Human identity recognition in aerial images. In *Proc. of CVPR*, pages 709–716. IEEE, 2010. 23

[157] J. Orwell, P. Remagnino, and G.A. Jones. Multi-camera colour tracking. In *Proc. of IEEE Workshop on Visual Surveillance (VS'99)*, pages 14–21. IEEE Comput. Soc, 1999. ISBN 0-7695-0037-4. 15, 29

[158] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, and Alan F. Smeaton. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011. 36

[159] R.E.A.C. Paley. On orthogonal matrices. *J. Math. Phys.*, 12:311–320, 1933. 117

[160] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita. ViSE: Visual Search Engine Using Multiple Networked Cameras. In *Proc. of ICPR*, page 1204, 2006. vii, 11, 17, 21, 29

[161] Unsang Park and Anil K. Jain. Face Matching and Retrieval Using Soft Biometrics. *IEEE Transactions. Inf. Forensics Security*, 5(3):406–415, September 2010. ISSN 1556-6013. 24

[162] Donovan H. Parks and Sidney S. Fels. Evaluation of background subtraction algorithms with post-processing. In *AVSS '08: Proceedings of the 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pages 192–199, Washington, DC, USA, 2008. IEEE Computer

Society. ISBN 978-0-7695-3341-4. doi: http://dx.doi.org/10.1109/AVSS.2008.19. 115

[163] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. of ICCV*, pages 261 –268, 29 2009-oct. 2 2009. 27

[164] A. G. Amitha Perera, Chukka Srinivas, Anthony Hoogs, Glen Brooksby, and Wensheng Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. *Proc. of CVPR*, 1:666–673, 2006. ISSN 1063-6919. doi: http://doi.ieeecomputersociety.org/10.1109/CVPR.2006.195. 31

[165] Valery A. Petrushin, Gang Wei, and Anatole V. Gershman. Multiple-camera people localization in an indoor environment. *Knowl. Inf. Syst.*, 10:229–241, August 2006. ISSN 0219-1377. 15

[166] PETS. Pets: Performance evaluation of tracking and surveillance, 2000–2009. http://www.cvg.cs.rdg.ac.uk/slides/pets.html. 39, 86, 91

[167] Thang V. Pham, Marcel Worring, and Arnold W.M. Smeulders. A Multi-Camera Visual Surveillance System for Tracking of Reoccurrences of People. In *Proc. of Int. Conf. on Distributed Smart Cameras*, pages 164–169. IEEE, September 2007. ISBN 978-1-4244-1353-9. 17, 21

[168] M. Piccardi. Background subtraction techniques: a review. In *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics*, volume 4, pages 3099 – 3104 vol.4, oct. 2004. doi: 10.1109/ICSMC.2004.1400815. 2, 115

[169] Gerard Pons-Moll, Laura Leal-Taixé, Tri Truong, and Bodo Rosenhahn. Efficient and robust shape matching for model based human motion capture. In *Proceedings of the 33rd international conference on Pattern recognition*, DAGM'11, pages 416–425, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-23122-3. URL http://dl.acm.org/citation.cfm?id=2039976.2040027. 33

[170] F. Porikli. Inter-camera color calibration by correlation model function. In *Proc. of ICIP*, volume 2, pages II – 133–6 vol.3, sept. 2003. 26

[171] W.K. Pratt, J. Kane, and H.C. Andrews. Hadamard transform image coding. *Proceedings of the IEEE*, 57(1):58–68, January 1969. 117, 119

[172] Cantata project. Video and image datasets index. online, 2008. URL http://www.multitel.be/cantata/. 38

[173] Bryan Prosser, Shaogang Gong, and Tao Xiang. Multi-camera Matching under Illumination Change Over Time. In *Proc. of Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Marseille, France, 2008. Andrea Cavallaro and Hamid Aghajan. 32

[174] Richard J. Radke. A survey of distributed computer vision algorithms. In *Aghajan (Eds.), Handbook of Ambient Intelligence and Smart Environments*. Springer, 2008. 18

[175] R.J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294 –307, march 2005. ISSN 1057-7149. doi: 10.1109/TIP.2004.838698. 115

[176] V. Reddy, C. Sanderson, and B.C. Lovell. An efficient and robust sequential algorithm for background estimation in video surveillance. In *Proc. of IEEE Int'l Conference on Image Processing*, pages 1109–1112, 2009. xi, 115, 116, 119, 121, 124, 125, 126

[177] Daniel Reid and Mark Nixon. Using comparative human descriptions for soft biometrics. In *The first International Joint Conference on Biometrics*, October 2011. URL http://eprints.soton.ac.uk/272922/. Event Dates: 11-13 October 2011. 24

[178] E. Roullot. A unifying framework for color image calibration. *Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on*, pages 97–100, 2008. doi: 10.1109/IWSSIP.2008.4604376. 25

[179] D. Russell and Shaogang Gong. A highly efficient block-based dynamic background model. In *Proceedings of IEEE Conference on Advanced Video*

and Signal Based Surveillance. AVSS 2005., pages 417 – 422, sept. 2005. doi: 10.1109/AVSS.2005.1577305. 115

[180] M. Salzmann and R. Urtasun. Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 647 –654, june 2010. doi: 10.1109/CVPR.2010.5540155. 33

[181] 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. 24

[182] Riccardo Satta, Giorgio Fumera, and Fabio Roli. Fast person re-identification based on dissimilarity representations. *Pattern Recognition Letters, Special Issue on Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context*, 33:1838–1848, 10/2012 2012. 27

[183] Philipp Schügerl, Robert Sorschag, Werner Bailer, and Georg Thallinger. Object re-detection using SIFT and MPEG-7 color descriptors. *Lecture Notes In Computer Science*, pages 305–314, 2007. ISSN 0302-9743. 17

[184] W.R. Schwartz and L.S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proc. of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009. vii, 23, 37, 39

[185] 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 *Proc. of CVPR*, pages 1297 –1304, june 2011. doi: 10.1109/CVPR.2011.5995316. 33, 77

[186] Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. 3d ellipsoid fitting for multi-view gait recognition. In *Proc. of AVSS*, pages 355 –360, 30 2011-sept. 2 2011. doi: 10.1109/AVSS.2011.6027350. 24

[187] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ISBN 1-59593-495-2. doi: http://doi.acm.org/10.1145/1178677.1178722. 39

[188] Bi Song, Ting-Yueh Jeng, Elliot Staudt, and Amit K. Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. In *Proc. of ECCV*, ECCV'10, pages 605–619, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15548-0, 978-3-642-15548-2. URL http://dl.acm.org/citation.cfm?id=1886063.1886109. 12

[189] H. Sorensen, D. Jones, M. Heideman, and C. Burrus. Real-valued fast fourier transform algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(6):849 – 863, jun 1987. ISSN 0096-3518. 119

[190] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. of CVPR*, pages 246–252, 1999. 115

[191] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions. on Pattern Analysis and Machine Intelligence*, 22(8):747–757, August 2000. 107

[192] Tatsuya Tanaka, Atsushi Shimada, Daisaku Arita, and Rin ichiro Taniguchi. A fast algorithm for adaptive background model construction using parzen density estimation. In *AVSS '07: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 528–533, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 978-1-4244-1695-0. doi: http://dx.doi.org/10.1109/AVSS.2007.4425366. 115

[193] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew W. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pages 103–110, 2012. 33, 77

[194] Luis F Teixeira and Luis Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 30(2):157–167, 2009. 27, 32

[195] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE J. of Robotics and Automation*, 3(4):323–344, 1987. 107

[196] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *T-PAMI*, 30(10):1713 –1727, oct. 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.75. 2, 48, 50, 102, 103

[197] Á. Utasi and Cs. Benedek. Multi-camera people localization and height estimation using multiple birth-and-death dynamics. In *Workshop on Visual Surveillance*, 2010. x, 107, 108, 112

[198] A. Utsumi and N. Tetsutani. Human tracking using multiple-camera-based head appearance modeling. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 657–662. IEEE, 2004. ISBN 0-7695-2122-3. 12

[199] K. van de Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *Proc. of CVPR*, pages 1 –8, june 2008. 23, 26

[200] 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)*, Snowbird, Utah, December 2009. 34

[201] C. Velardo and J. Dugelay. Weight estimation from visual body appearance. In *2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS),*, pages 1 –6, sept. 2010. 24

[202] Roberto Vezzani and Rita Cucchiara. Video surveillance online repository (visor): an integrated framework. *Multimedia Tools and Applications*, 50 (2):359–380, November 2010. 36, 38, 126

[203] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. Pathnodes Integration of Standalone Particle Filters for People Tracking on Distributed Surveillance Systems. In *Proc. of IEEE Int. Conf. on Image Analysis and Processing*, pages 404–413, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04145-7. 13, 20, 60

[204] Roberto Vezzani, Costantino Grana, and Rita Cucchiara. Probabilistic people tracking with appearance models and occlusion classification: The ad-hoc system. *Pattern Recognition Letters*, 32(6):867–877, April 2011. 58

[205] Paul Viola, John C. Platt, and Cha Zhang. Multiple instance boosting for object detection. In *In NIPS 18*, pages 1419–1426. MIT Press, 2006. 27

[206] Xuan-he Wang and Ji-lin Liu. Tracking multiple people under occlusion and across cameras using probabilistic models. *Journal of Zhejiang University SCIENCE A*, 10(7):985–996, July 2009. ISSN 1673-565X. 13, 23

[207] Michael Weber and Martin Bauml. Part-based clothing segmentation for person retrieval. In *Proc. of AVSS*, pages 361 –366, 30 2011-sept. 2 2011. 16, 23

[208] Bo Yang, Chang Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a crf model. In *Proc. of CVPR*, pages 1233 –1240, june 2011. 27, 28

[209] Jie Yang, Xiaojin Zhu, Ralph Gross, John Kominek, Yue Pan, and Alex Waibel. Multimodal people ID for a multimedia meeting browser. In *Proc. of Int. ACM Multimedia Conference*, page 159, 1999. 12

[210] X.-S. Yang and S. Deb. Engineering optimisation by cuckoo search. *Int. J. Mathematical Modelling and Numerical Optimisation*, 1:330–343, 2010. 82

[211] Meng Yao and Huchuan Lu. Human body segmentation in a static image with multiscale superpixels. In *Awareness Science and Technology (iCAST), 2011 3rd International Conference on*, pages 32 –35, sept. 2011. doi: 10.1109/ICAwST.2011.6163091. 48, 50

[212] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/1177352.1177355. 2, 31

[213] K. Yoon, D. Harwood, and L. Davis. Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation*, 17(3):605–622, June 2006. ISSN 10473203. 23

[214] Yang Yu, David Harwood, Kyongil Yoon, and Larry S. Davis. Human appearance modeling for matching across video sequences. *Machine Vision and Applications*, 18(3-4):139–149, April 2007. ISSN 0932-8092. 14, 21

[215] W. Zajdel, Z. Zivkovic, and B.J.A. Krose. Keeping track of humans: Have i seen this person before? In *Proc. of IEEE Int. Conf. on Robotics and Automation, ICRA 2005.*, pages 2081 – 2086, April 2005. doi: 10.1109/ROBOT.2005.1570420. 2, 15

[216] Li Zhang, Yuan Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Proc. of CVPR*, pages 1 –8, june 2008. 31

[217] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proc. of BMVC*, 2009. 17, 23

[218] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proc. of CVPR*, pages 649–656, 2011. 16, 23, 28, 33, 36, 78, 85

[219] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Re-identification by relative distance comparison. *T-PAMI*, 99(PrePrints):1–1, 2012. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.138. 18, 28, 33

[220] Q Zhou and J Aggarwal. Object tracking in an outdoor environment using fusion of features and cameras. *Image and Vision Computing*, 24(11):1244–1255, November 2006. ISSN 02628856. 14

[221] Yingbo Zhou and A. Kumar. Human identification using palm-vein images. *IEEE Transactions. Inf. Forensics Security*, 6(4):1259 –1274, Dec. 2011. ISSN 1556-6013. doi: 10.1109/TIFS.2011.2158423. 24