

People Reidentification in Surveillance and Forensics: A Survey

ROBERTO VEZZANI, DAVIDE BALTIERI, and RITA CUCCHIARA, University of Modena and Reggio Emilia

The field of surveillance and forensics research is currently shifting focus and is now showing an ever increasing interest in the task of people reidentification. This is the task of assigning the same identifier to all instances of a particular individual captured in a series of images or videos, even after the occurrence of significant gaps over time or space. People reidentification can be a useful tool for people analysis in security as a data association method for long-term tracking in surveillance. However, current identification techniques being utilized present many difficulties and shortcomings. For instance, they rely solely on the exploitation of visual cues such as color, texture, and the object's shape. Despite the many advances in this field, reidentification is still an open problem. This survey aims to tackle all the issues and challenging aspects of people reidentification while simultaneously describing the previously proposed solutions for the encountered problems. This begins with the first attempts of holistic descriptors and progresses to the more recently adopted 2D and 3D model-based approaches. The survey also includes an exhaustive treatise of all the aspects of people reidentification, including available datasets, evaluation metrics, and benchmarking.

Categories and Subject Descriptors: A.1 [General Literature]: Introductory and Survey; I.4.7 [Image Processing and Computer Vision]: Feature Measurement; I.5.4 [Pattern Recognition]: Applications

General Terms: Documentation, Algorithms

Additional Key Words and Phrases: People reidentification, computer vision, multimedia surveillance

ACM Reference Format:

Vezzani, R., Baltieri, D., Cucchiara, R. 2013. People reidentification in surveillance and forensics: A survey. *ACM Comput. Surv.* 46, 2, Article 29 (November 2013), 37 pages.

DOI: <http://dx.doi.org/10.1145/2543581.2543596>

1. INTRODUCTION

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 applications 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.

This work was mainly carried out within the project THIS (JLS/2009/CIPS/AG/C1-028), with the support of the Prevention, Preparedness and Consequence Management of Terrorism and Other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security. The work was also partially supported by EU POR-FESR Emilia Romagna funds for the research activity in surveillance at the SOFTECH-ICT Center of Modena's Technopole, Italy.

Authors' address: R. Vezzani, D. Baltieri, and R. Cucchiara, Dipartimento di Ingegneria "Enzo Ferrari," University of Modena and Reggio Emilia, Via Vignolese, 905/b 41125, Modena - Italy; emails: {roberto.vezzani, davide.baltieri, rita.cucchiara}@unimore.it.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2013 ACM 0360-0300/2013/11-ART29 \$15.00

DOI: <http://dx.doi.org/10.1145/2543581.2543596>

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 [Piccardi 2004]; people tracking, which exploits time coherency to follow the same object/person along time and space [Yilmaz et al. 2006]; and people detection by appearance, which adopts machine learning techniques for object (and in particular human and pedestrian) detection [Dalal et al. 2006; Gualdi et al. 2011; Tuzel et al. 2008]. 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. [2007].

The most recent topic to generate interest in surveillance is **people reidentification** in image and video archives. People reidentification 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 an image or a video.

With this premise, people reidentification aims to answer questions such as “Where have I seen this person before?” [Zajdel et al. 2005] or “Where has he gone after being caught on this surveillance camera?” Reidentification works on the exterior appearance usually acquired by noisy cameras, which makes it impossible to extract and associate precise measurements such as biometric features. 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 [Frontex 2011] 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 predefined unique object.
- Verify*: Given prior knowledge on the object, can its presence/position be confirmed?

In agreement with the EUROSUR definition, reidentification lies in between identification and recognition. It can be associated with the *identification* task, assuming that the goal of reidentification is to match different observations of people 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. Reidentification 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.

Reidentification 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 and 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 noncollaborative targets and when biometric recognition is not feasible.

Thus, people reidentification 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.

Table I. Continuity Constraints Imposed by People Tracking, Reidentification, and Biometric Recognition

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

Most methodologies of people reidentification 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” [Forsyth and Ponce 2002]. On the contrary, biometry is devoted to finding the exact identity of each piece of evidence. Different conditions and hypotheses on the spatiotemporal 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 hypotheses of constancy and they try to solve additional challenges such as illumination changes, noise, occlusions, and so on.

Different from people tracking, the reidentification task aims to match people instances during a time delay and/or a change in point of view. This invalidates the first two continuity constraints, while the global appearance, in addition to the biometric profile, is preserved (see Table I).

Thus, reidentification 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, reidentification 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, reidentification, and biometry are slowly fading, leaving behind a plethora of methods that fall in between the two classes. Examples include soft biometry [Jain et al. 2004] and tracking algorithms by data association. Some tracking algorithms designed to handle occlusion issues relax temporal continuity constraints and resemble reidentification in the way that they share similar methodologies. In addition, the recent “tracking-by-detection” approach [Andriluka et al. 2008] 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 (position in the 2D camera image plane and point of view). The

“tracking-by-identification” approach proposed by Rios-Cabrera et al. [2011] is another borderline case.

Although the boundaries between reidentification and tracking cannot be easily defined, this survey will present the characterizing aspects and main issues concerning the methods specifically designed for people reidentification. The information on the methods provided by this survey could be applied as support for short- or long-term tracking in surveillance or as a tool for people image retrieval in typical forensics applications.

Despite the potential impact of this study, it should be pointed out that research in this field is still at an early stage. This study is part of a conglomeration of current proposals that are still investigating a very large solution space. Reidentification is a difficult, laborious, and somewhat ambiguous task since visual cues characterizing a person’s aspect can be poorly distinctive.

The goal of this article is to discuss several aspects of the reidentification process, which should leave us with a more complete view of the challenges this field is currently faced with and the possible ways to overcome them. We provide a conceptualization of the reidentification task by describing in detail the different dimensions of the current problems and previously proposed solutions. The main contribution of this work is a very large survey of existing proposals and the conceptualization of the aspects and the dimensions of the solution space. In particular, the focus of the discussion concerns the feature space and its connections to human body models, ranging from generic holistic descriptors to specific models based on 2D and 3D representation.

The article is structured as follows: Section 2 presents a multidimensional taxonomy of the available reidentification techniques. Section 3 provides a structured review of the topic’s literature. Section 4 illustrates the datasets and metrics available for evaluation and testing of reidentification performances. Finally, Section 5 provides a general discussion on the research in this field where new promising directions are highlighted.

2. REIDENTIFICATION: A MULTIDIMENSIONAL OVERVIEW

Research in surveillance and people analysis for security has been thoroughly focused on people reidentification 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, we propose a multidimensional taxonomy of the problem. Instead of adopting a hierarchical taxonomy where classification criteria are placed at each level (such as in the work by Aggarwal and Cai [1999]), we propose a multidimensional space as illustrated in Figure 1. Reidentification 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.

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 reidentification 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 reidentification 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

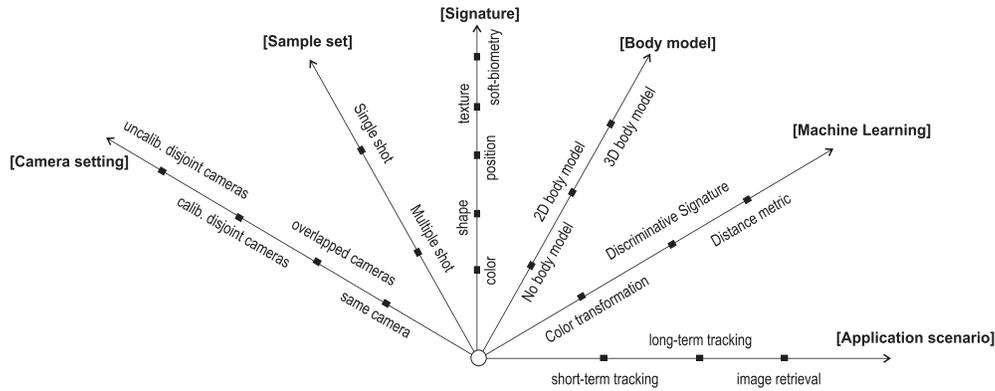


Fig. 1. Multidimensional taxonomy for people reidentification algorithms.

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 data availability, the reidentification process may handle 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, reidentification 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. Reidentification 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**. Even if the body model could be considered another signature feature, a specific dimension can better highlight its crucial importance to categorize methods and approaches. Extracted features can be computed on different spatial levels, varying from holistic features 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 new 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

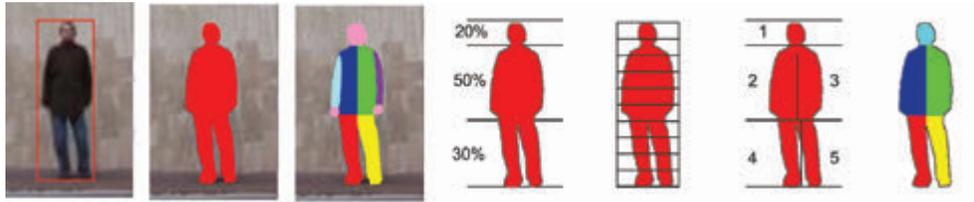


Fig. 2. From left to right: region of interest: bounding box by people detection, complete pixel-wise silhouette from foreground segmentation, face/body part segmentation and classification. Examples of body models: three horizontal fixed slices [Albu et al. 2006; Park et al. 2006; Lantagne et al. 2003], 10 slices [Bird et al. 2005], symmetry-based parts [Farenzena et al. 2010], five human body parts [Bak et al. 2010].

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 Figure 2).

The fifth dimension corresponds to the exploitation of **machine learning** algorithms during the reidentification 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.

The specific **application scenario** is the case where the reidentification 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.

Figure 1 summarizes the six dimensions that 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.

3. FIFTEEN YEARS OF RESEARCH IN REIDENTIFICATION

Since the paper by Cai and Aggarwal [1996] first described an attempt to follow the same person using a multicamera scenario, roughly 100 papers on the subject have been proposed, which have been classified and summarized (see Table II). 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.

3.1. Camera Setting

Apart from some initial experiments of people reidentification as a particular case of shape classification (e.g., the seminal work by Cai and Aggarwal [1998]), 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 [2006] presented a statistical method to learn the environment's topology using a large amount of tracking data; Calderara et al. [2008a] and Khan and Shah [2003] 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 [2008].

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

Table II. Examples of Reidentification Methods Classified with the Multidimensional Taxonomy (grouped by main application scenario and in chronological order)

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
Reidentification FOR TRACKING					
[Albiol et al. 2012]	Uncal. Disj. Hi-res. Indoor	Multiple	Color [RGB] <i>Bodyprints</i>	3D - 90 stripes	p.v.
[Mazzon et al. 2012]	Uncal. Disj. Med-res. Indoor	Single	Color [RGB, YCbCr, HSV], Texture <i>Color, Schmid and Gabor texture features</i>	none	i-LIDS (MCTS)
[Brendel et al. 2011]	Same Med-res.	Single	Color [RGB], Shape <i>RGB histogram, gradients, and optical fFlow</i>	none	ETHZ Central, TUD crossing, i-Lids, UBC Hockey, ETHZ Soccer
[Madrigal and Hayet 2011]	Overlapping Med-res. Outdoor	Multiple	Color [HSV], Position <i>Color histogram and feet position</i>	none	PETS2009, Caviar
[Jungling and Arens 2011]	Uncal. Disj. Med-res. Indoor	Multiple	Texture <i>SIFT and ISM</i>	none	CasiaA
[Conte et al. 2011]	Uncal. Disj. Med-res. Outdoor	Multiple	Color [RGB] <i>Appearance mask</i>	none	PETS2010
[Kuo and Nevatia 2011]	Same Med-res. In-door/Outdoor	Single	Color [RGB], Shape, Texture <i>HOG and covariance</i>	none	Caviar, Trecvid08 and ETH
[Dantcheva et al. 2011]	Same Uncal. Disj. Calib. Disj. Hi-res.	Single	Color [HSV] <i>Bag of soft biometric features: eye, hair, skin and clothes color; eyeglasses, beard and mustache presence, weight</i>	none	p.v., Caviar, ViPER
[Cong et al. 2010b]	Uncal. Disj. Hi-res. Indoor	Multiple	Color [RGB] <i>Color-position histogram</i>	2D - LTH with fixed slices	p.v.
[Alahi et al. 2010]	Calib. Disj. Lo-res. In-door/Outdoor	Single	Color [several], Shape, Texture, Position	2D - grids	Viper, p.v.
[Kuo et al. 2010]	Same Calib. Disj. Med-res.	Multiple	Color [RGB], Shape, Texture <i>RGB histogram, covariance matrix, HOG</i>	none	CAVIAR, TRECVID08
[Lian et al. 2010]	Overlapping Hi-res. Outdoor	Single	Position <i>Adaptive homographies at different levels</i>	none	p.v.
[Song et al. 2010]	Calib. Disj. Med-res.	Single	Shape <i>RGB histogram</i>	none	CAVIAR, VideoWeb

(Continued)

Table II. Continued

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
Reidentification FOR TRACKING					
[Jungling and Arens 2010]	Same Med-res. Indoor	Multiple	Texture <i>SIFT on infrared images</i>	none	Casia Infrared dataset
[Vezzani et al. 2009]	Calib. Disj. Lo-res. Outdoor	Single	Color [RGB], Position <i>Color histogram and position</i>	none	ViSOR
[Wang and Liu 2009]	Overlapping Lo-res. Outdoor	Single	Color [RGB] <i>Color histogram</i>	2D - body parts	p.v.
[Denman et al. 2009]	Same Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB] <i>Height of the LTH parts and color histograms</i>	2D - LTH	PETS2006
[Anjum and Cavallaro 2009]	Overlapping Hi-res. Indoor	Single	Color [n.a.], Position <i>Position on the ground plane and image appearance</i>	none	Sport videos
[Li et al. 2009a]	Overlapping Hi-res. Indoor	Single	Position <i>Vertical axis and homography projections</i>	none	p.v.
[Calderara et al. 2008a]	Overlapping Lo-res. Outdoor	Single	Shape, Position <i>Feet and head position, vertical axis</i>	none	ViSOR
[Jing-Ying et al. 2008]	Overlapping Lo-res. Outdoor	Single	Position <i>Feet position</i>	none	PETS2001
[Javed et al. 2008]	Calib. Disj. Hi-res. Outdoor	Multiple	Color [RGB], Position <i>Feet position and brightness transfer function</i>	none	Online cameras
[Hyodo et al. 2008]	Overlapping Med-res. Outdoor	Single	Position <i>Feet position and body height</i>	2D rectangular with fixed height	p.v.
[Hu et al. 2008]	Same Med-res. Indoor	Single	Color [RGB], Texture <i>SIFT and color autocorrelogram</i>	none	Caviar
[Gray and Tao 2008]	Uncal. Disj. Lo-res. Outdoor	Single	Color [RGB, HSV, YCbCr], Texture <i>Mean values</i>	rectangular stripes	ViPER
[Chen et al. 2008]	Uncal. Disj. Calib. Disj. Med-res.	Single	Color [RGB] <i>RGB histogram + calibrazione + camera network topology</i>	none	p.v.
[Colombo et al. 2008a]	Uncal. Disj. Med-res. Indoor	Single	Color [YCbCr], Texture <i>Mean color, cov. matrix and others</i>	none	Torino metro station

(Continued)

Table II. Continued

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
Reidentification FOR TRACKING					
[Yu et al. 2007]	Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB], Shape <i>Color path-length profile</i>	none	Honeywell
[Madden et al. 2007]	Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB] <i>MCSHR: color clusters</i>	none	p.v.
[Albu et al. 2006]	Uncal. Disj. Med-res. Indoor	Single	Color [n.a.], Texture <i>Spatial and spectral distribution of dominant colours</i>	2D - LTH	Online cameras
[Loke et al. 2006]	Calib. Disj. Outdoor	Single	Shape, Position <i>Width, height, motion, and position</i>	none	p.v.
[Hu et al. 2006]	Overlapping Med-res. Outdoor	Single	Position <i>Principal axis</i>	none	PETS2001, p.v.
[Berclaz et al. 2006]	Overlapping Med-res.	Single	Color [RGB] <i>RGB histogram</i>	none	p.v.
[Zhou and Aggarwal 2006]	Overlapping Med-res. Outdoor	Single	Position <i>Feet position</i>	none	PETS2001
[Gilbert and Bowden 2006]	Calib. Disj. Med-res. Indoor	Single	Color [RGB] <i>RGB histogram</i>	none	p.v.
[Gandhi and Trivedi 2006]	Uncal. Disj. Med-res. Indoor	Multiple	Color [n.a.] <i>Appearance map</i>	3D - Cylinder - PAM	p.v.
[Petrushin et al. 2006]	Uncal. Disj. Med-res. Indoor	Single	Color [RGB] <i>Color histogram</i>	none	p.v.
[Zajdel et al. 2005]	Same Med-res. Indoor	Single	Color [RGB] <i>Mean color</i>	2D - LTH with fixed slices	Icra05
[Utsumi and Tetsutani 2004]	Same Indoor/ Outdoor	Single	Color [n.a.], Texture <i>Head appearance</i>	none	p.v.
[Khan and Shah 2003]	Overlapping Med-res. Outdoor	Single	Position <i>Feet position</i>	none	PETS2001, p.v.
[Black et al. 2002]	Overlapping Calib. Disj. Outdoor	Multiple	Position <i>Feet position</i>	none	PETS2001, p.v.
[Chang and Gong 2001]	Overlapping Lo-res. Indoor	Single	Color [HSV], Position <i>Feet and head position, color GMM, height</i>	none	Online cameras
[Krumm et al. 2000]	Same Med-res. Indoor	Multiple	Color [RGB] <i>Color histogram for each region of the scene</i>	none	Online cameras

(Continued)

Table II. Continued

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
[Kettner and Zabih 1999]	Calib. Disj. Lo-res. Indoor	Single	Color [HSV], Position <i>Color, speed and spatiotemporal camera model</i>	none	p.v.
[Cai and Aggarwal 1999]	Overlapping Lo-res. Indoor	Single	Position <i>Feet position and mean intensity</i>	none	Online cameras
[Yang et al. 1999]	Same Indoor	Single	Color [rg] <i>Color histogram, face and voice</i>	none	p.v.
[Orwell et al. 1999]	Same Uncal. Disj. Med-res. Outdoor	Single	Color [YUV] <i>Color histogram</i>	none	p.v.
Reidentification FOR RETRIEVAL					
[Zheng et al. 2012]	Uncal. Disj. Med-res. In- door/Outdoor	Single	Color [RGB, HS,YCbCr], Texture <i>Color, Schmid and Gabor texture features</i>	2D - 6 stripes	i-LIDS (MCTS), ViPER, ETHZ
[Hirzer et al. 2012]	Uncal. Disj. Med-res. In- door/Outdoor	Single	Color [HSV, Lab], Texture <i>Mean color and LBP histogram</i>	2D - Grid	ViPER, ETHZ, Prid2011
[Bazzani et al. 2012]	Uncal. Disj. Med-res. In- door/Outdoor	Multiple	Color [HSV], Texture <i>Histogram plus epitome</i>	2D - LTH plus simmetry based vertical splits	i-LIDS, ETHZ, CAVIAR4REID
[Liu et al. 2012]	Uncal. Disj. Med-res. In- door/Outdoor	Single	Color [RGB, YCbCr,HSV], Texture <i>Color, Schmid and Gabor texture features</i>	2D - 6 stripes	i-LIDS (MCTS), ViPER
[Layne et al. 2012]	Uncal. Disj. Hi-res. In- door/Outdoor	Single	Color [] <i>High-level attributes</i>	none	i-LIDS (MCTS), ViPER, ETHZ
[Fischer et al. 2011]	Same Hi-res. Indoor	Single	Texture <i>DCT-based facial appearance</i>	2D - face	p.v.
[Baltieri et al. 2011c]	Uncal. Disj. Med-res. In- door/Outdoor	Multiple	Color [HSV] <i>Color histogram</i>	3D model	ViSOR, Sarc3D
[Weber and Bauml 2011]	Uncal. Disj. Hi-res. Indoor	Single	Color [RGB] <i>Color histogram</i>	2D - upper and lower part	p.v.
[Zheng et al. 2011]	Uncal. Disj. Med-res. In- door/Outdoor	Single	Color [RGB, YCbCr,HSV], Texture <i>Color, Schmid and Gabor texture features</i>	2D - 6 stripes	i-LIDS (MCTS), ViPER

(Continued)

Table II. Continued

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
Reidentification FOR RETRIEVAL					
[Dantcheva and Dugelay 2011]	Uncal. Disj. Hi-res. Indoor	Single	Color [HSV], Texture <i>Texture and color</i>	none	Feret
[Aziz et al. 2011]	Uncal. Disj. Hi-res. Indoor	Single	Texture <i>SIFT</i> , <i>SURF</i> , <i>Spin</i>	2D - LTH	p.v.
[Bauml and Stiefelhagen 2011]	Uncal. Disj. Lo-res. Indoor	Multiple	Texture <i>SIFT</i> , <i>SURF</i> , <i>SC</i> , <i>GLOH</i>	none	Caviar
[Bak et al. 2011]	Uncal. Disj. Med-res. Indoor	Multiple	Color [RGB], Shape <i>MRCG - mean</i> <i>Riemaniann</i> <i>covariance grid</i>	none	i-LIDS, ETHZ
[Farenzena et al. 2010]	Same Uncal. Disj. In- door/Outdoor	Multiple	Color [HSV], Shape, Texture <i>Weighted color</i> <i>histograms</i> , <i>MSCR</i> , <i>recurrent</i> <i>high structured</i> <i>patches</i>	2D - LTH plus symmetry- based vertical splits	ViPER, i-LIDS, ETHZ
[Metternich et al. 2010]	Uncal. Disj. Lo-res. Outdoor	Single	Color [Several], Shape <i>Color histogram</i>	none	ViPER
[Bak et al. 2010]	Uncal. Disj. Med-res. Indoor	Single	Color [RGB], Texture <i>Haar-based and</i> <i>DCD-based</i> <i>signature</i>	2D - body parts	i-LIDS (MCTS)
[Ali et al. 2010]	Uncal. Disj. Lo-res.	Single	Color [RGB, YCbCr,HSV], Texture <i>RGB</i> , <i>YCbCr</i> , <i>HSV</i> <i>histogram</i> , <i>Schmid and Gabor</i> <i>filters responses</i>	none	ViPER
[Prosser et al. 2010]	Uncal. Disj. Med-res. In- door/Outdoor	Single	Color [RGB, HS,YCbCr], Texture <i>Color</i> , <i>Schmid and</i> <i>Gabor texture</i> <i>features</i>	2D - 6 stripes	i-LIDS (MCTS), ViPER
[Liu and Yang 2009]	Uncal. Disj. Med-res. Indoor	Multiple	Color [YCbCr], Texture <i>Bag of siftch -</i> <i>color SIFT</i>	none	Caviar, ViPER
[Zheng et al. 2009]	Uncal. Disj. Hi-res. Indoor	Single	Color [RGB], Texture <i>CRRRO</i> <i>descriptor: SIFT</i> <i>and color</i>	none	i-LIDS (MCTS)
[Monari et al. 2009]	Overlapping Uncal. Disj. Lo-res. Indoor	Single	Color [CIEluv] <i>Mean color</i>	2D - LTH with fixed slices	Online cameras

(Continued)

Table II. Continued

Ref.	CAMERA SETTING	SAMPLE SET	SIGNATURE	BODY MODEL	DATASET
Reidentification FOR RETRIEVAL					
[de Oliveira and Pio 2009]	Uncal. Disj. Lo-res. Indoor	Single	Color [modified HSV], Texture <i>SURF</i>	none	Caviar, Weizmann
[Hamdoun et al. 2008]	Uncal. Disj. Lo-res. Indoor	Multiple	Texture <i>Set of SURF-like descriptors</i>	none	Caviar
[Lin and Davis 2008]	Uncal. Disj. Calib. Disj. Med-res. Indoor	Multiple	Color [rgb] <i>Color rank</i>	none	Honeywell dataset
[Pham et al. 2007]	Uncal. Disj. Med-res. Outdoor	Single	Color [RGB] <i>Weighted color histogram</i>	2D - standard human mask	Online cameras
[Schügerl et al. 2007]	Uncal. Disj. Indoor/ Outdoor	Single	Color [YCbCr], Texture <i>SIFT and MPEG7 color layout</i>	none	TRECVID
[Park et al. 2006]	Overlapping Uncal. Disj. Calib. Disj. Med-res. Outdoor	Single	Color [HSV], Shape, Position <i>Mean color; feet position, height, bodybuild ratios</i>	2D - LTH with fixed slices	p.v.
[Gheissari et al. 2006]	Uncal. Disj. Med-res. Outdoor	Single	Color [modified HSV, RGB], Texture <i>Appearance map</i>	2D - spatio-temporal appearance model	p.v.
[Bird et al. 2005]	Same Med-res. Outdoor	Single	Color [HSL] <i>Median color</i>	2D - 10 horizontal slices	p.v. - bus stop
[Nakajima et al. 2003]	Uncal. Disj. Lo-res. Indoor	Single	Color [RGB, rgb], Shape <i>Color histogram and shape features</i>	none	p.v.
[Lantagne et al. 2003]	Uncal. Disj. Med-res. Indoor	Single	Color [HSV], Texture <i>Dominant color; histograms, edge energy</i>	2D - LTH with fixed slices	p.v.

Same camera: In this setting, the goal of reidentification is to be able to identify the same individual repeatedly using the same camera after he or she is initially detected [Bird et al. 2005; Fischer et al. 2011]. 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 viewpoint 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. [2005] describes an application in public transportation areas, while the work by Jungling and Arens [2010] designs reidentification 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. [2011].

Overlapping cameras: In this scenario, reidentification can be considered as a part of a long-term tracking process over enlarged fields of view. This problem is also called *consistent labeling* [Cai and Aggarwal 1998; Hu et al. 2006; Calderara et al. 2008a; Khan and Shah 2003]. 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 reidentification process can even operate in crowded scenarios where multiple individuals are occluding one another [Khan and Shah 2009]. The first work on people matching and reidentification was proposed by Cai and Aggarwal [1996] 15 years ago. An initial manual or automatic [Khan and Shah 2003; Calderara et al. 2008a] estimation of the homography matrix allows the reidentification problem to be reduced to geometric-based matching of the feet coordinates. In addition, epipolar constraints can be exploited [Hartley and Zisserman 2004; Calderara et al. 2008b]. A plethora of systems based on this assumption and relationship have been proposed [Cai and Aggarwal 1998; Lee et al. 2000; Ellis and Black 2003; Hu et al. 2006; Hyodo et al. 2008; Khan and Shah 2009; Monari et al. 2009; Anjum and Cavallaro 2009].

Calibrated disjoint cameras: This setting is oriented toward large-area surveillance with known camera layout. Even if the cameras' fields of view are nonoverlapping, some geometrical information can still be useful [Kettner and Zabih 1999; Black et al. 2002; Javed et al. 2008; Alahi et al. 2010]. By using a homography transformation to obtain feet position on a common ground plane [Lee et al. 2000], the temporal gap between two corresponding views can be bridged [Makris et al. 2004]. This is made possible by means of predictive filters such as the Kalman filter [Ellis and Black 2003] or the particle filter [Vezzani et al. 2009]. 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. [2012].

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, and so forth, featuring nonhomogeneous capabilities and technologies. In this case the reidentification task is sometimes referred to as *reacquisition*, as suggested by Cong et al. [2010a]. 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 3.4 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. Reidentification by aspect similarity can be considered a form of soft biometry for people identification that 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 [Gheissari et al. 2006; Cong et al. 2010b]. As illustrated in Table II, the majority of works submitted in recent years fall into this category and address reidentification without any setting limitations.

Additional aspects of the camera setting, such as the point of view (viewing angle), image resolution (number of pixels), and image quality (compression, noise, etc.) strongly affect the reidentification 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, 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.

3.2. Sample Cardinality

The reidentification 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 [Bazzani et al. 2012]. 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., [Lantagne et al. 2003; Park et al. 2006; Albu et al. 2006; Pham et al. 2007; Bak et al. 2010]), the following information 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. [2010b] selected 10 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. [2007] 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. [2010]. 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 [Cai and Aggarwal 1999]. The best view selection is also effective in the presence of occlusions [Khan and Shah 2009].

Set of specialized signatures: A set of signatures is computed and stored for each person. Different 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. [2000] computes a different training signature for each (discretized) person's position in the scene and each time considers the subset of appearances extracted from 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., [Hamdoun et al. 2008; Jungling and Arens 2010; Liu and Yang 2009]).

Body-model-based signature: The final signature directly integrates more contributions (e.g., [Gandhi and Trivedi 2006; Baltieri et al. 2010; Cheng et al. 2011]). For example, the PAM appearance map developed by Gandhi and Trivedi [2006] and the SARC3D model defined by Baltieri et al. [2011c] are obtained by updating the visible part of the signature.

3.3. Signature

As with most pattern recognition problems, reidentification efficacy is directly affected by the type of adopted signature. Works proposed until now have exploited different

features, which can be grouped approximately by color, shape, position, texture, and soft biometry. Recently, Doretto et al. [2011] 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 limit their usage and more generic features are required.

Color: Even if it depends on external illumination, camera technology, and setting, *color* is the feature exploited the most. Plain means, histograms [Nakajima et al. 2003; Madden et al. 2007; Wang and Liu 2009; Weber and Bauml 2011], and Gaussian models [Colombo et al. 2008a] 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 II. In some works, more color descriptors are mixed or compared, such as that found in the works of van de Sande et al. [2008], Gray and Tao [2008], Prosser et al. [2010], and Zheng et al. [2011].

Shape: features including width, height, width/height ratio [Huang and Russell 1998], vertical axis [Calderara et al. 2008a; Hu et al. 2006], moment invariants [Mindru et al. 2004], and contours [Yoon et al. 2006] 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 [Khan and Shah 2003; Hu et al. 2006; Calderara et al. 2008a]. Different from tracking algorithms, the position is not used to match detections extracted in subsequent frames of the same camera by means of the temporal continuity described in Section 1.

Texture: Covariance matrices [Bak et al. 2010], SIFT [Hu et al. 2008; Zheng et al. 2009], and SURF [Hamdoun et al. 2008] descriptors are some examples of texture-based features (see Figure 3). More recently, HoG-like descriptors [Oreifej et al. 2010; Schwartz and Davis 2009] and LBP features [Hirzer et al. 2012] have been proposed.

Mixed descriptors: Combinations of features have been evaluated with the goal of integrating color, shape, and texture contributions [Alahi et al. 2010; Farenzena et al. 2010; Gray and Tao 2008; Metternich et al. 2010] in the same signature. de Oliveira and Pio [2009] were able to improve the base SURF descriptors using HSV color information to create an integrated color and texture feature set. Ali et al. [2010] computed Schmidt and Gabor texture filters on different color spaces. Kang et al. [2005] 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 [2005] are a generalization of histograms that includes higher-order spatial moments. For example, the second-order spatiogram contains the spatial mean and covariance for each histogram bin. A detailed list of adopted features is illustrated in Table II.

Soft biometry: Iris scanning [Miyazawa et al. 2008], Palm-Vein images [Zhou and Kumar 2011], fingerprints, hand appearance [Dutagaci et al. 2008], and other hard biometric identifiers [Delac and Grgic 2004] are expressly excluded from this survey, but intermediate soft biometric features [Jain et al. 2004] such as gait [Sivapalan et al. 2011; Havasi et al. 2005], facial features [Dantcheva and Dugelay 2011; Park and Jain 2010; Fischer et al. 2011], body size [Denman et al. 2009], and body weight [Velardo and Dugelay 2010] are included. These features can be effectively analyzed for

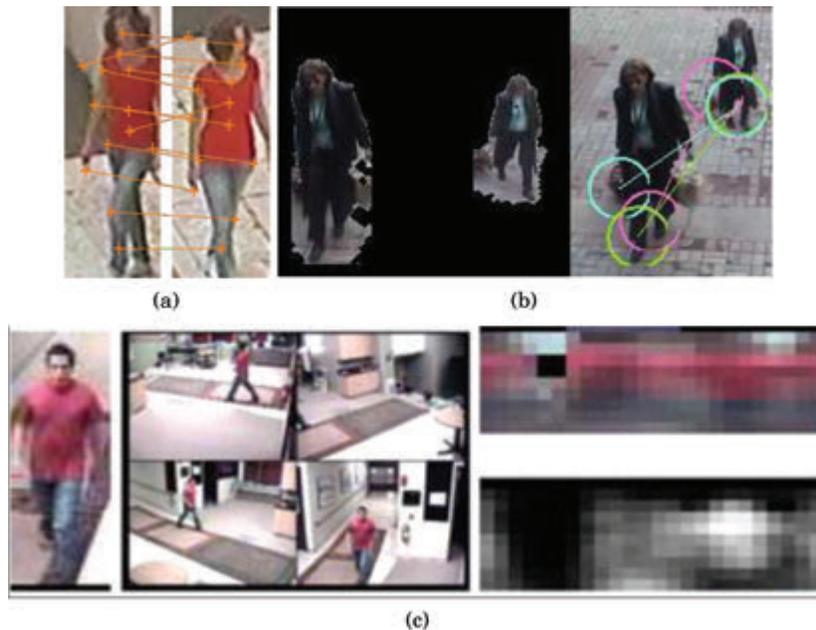


Fig. 3. Examples of proposed solutions: (a) SPIN, SURF, and SIFT local features; (b) interest operator matching in 2D body models [Doretto et al. 2011] (©Springer 2011); (c) 3D panoramic map with overlapped cameras [Gandhi and Trivedi 2006] (©IEEE 2006).

reidentification if the image resolution is sufficiently high [Dantcheva et al. 2010] or if an RGBD sensor is available [Barbosa et al. 2012]. 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 [Reid and Nixon 2011; Layne et al. 2012; Satta et al. 2012b].

An overview of the general topic of soft biometry and a new refined definition of the field has been provided by Dantcheva et al. [2011], 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 3.4). Following the aforementioned, recent proposals have shown no desire for a manual feature selection and provide the aforementioned features to the learning module [Hirzer et al. 2012]. Despite the fact that reidentification still performs with some level of mediocrity, great improvements could be made with the incorporation of new view invariant descriptors.

3.4. MACHINE LEARNING

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

Color transformation: Under the conditions of different camera types or where the illumination is not uniform, people matching using color-based signatures can



Fig. 4. A common problem of multicamera systems: (a) different views have different colors; (b) patterns used for the color calibration.

be affected by systematic variations in the input signals (see Figure 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 Figure 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 [Roullot 2008] as well as more complex nonlinear approaches [Gijzen et al. 2011] have been implemented. Despite dependence on a large number of parameters, Javed et al. [2008] 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. [2004] used a nonuniform quantization of the HSV color phase to improve illumination invariance, while Bowden and KaewTraKulPong [2005] and Gilbert and Bowden [2006] exploited the “Consensus - Color Conversion of the Munsell color space” (CCCM), a coarse quantization based on human perception. Porikli [2003] adopted a nonlinear 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. [2008a] 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. [2010] and van de Sande et al. [2008] present us with a set of descriptors and color spaces that appear invariant to the illumination conditions. When illumination changes are solely responsible for the incoherence of the colors among cameras, the brightness transfer function can be an invaluable tool learned and applied initially, as suggested by Porikli [2003] and improved by Javed and Shafique [2005] and Gilbert and Bowden [2006].

Discriminative signature: 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. Different machine-based learning 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. [2010], Babenko et al. [2009], and Mei and Ling [2011] create and update the object model using online learning algorithms that do not require any previous training. Conversely, Grabner et al. [2010] and Pellegrini et al. [2009] 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 [Yang et al. 2011] could be applied. Teixeira and Corte-Real [2009] introduced

an online learning step using a bag-of-features model based on SIFT descriptors. Similarly, Babenko et al. [2009] proposed creating the appearance model of each person using a Multiple Instance Learning (MIL) algorithm derived from MIL-Boost by Viola et al. [2006]. The main drawback of these techniques is the requirement of multiple source images to adopt a specific “class” that corresponds to a specific person in reidentification. Another inconvenience is that offline computations do not usually permit a fast automatic update mechanism when new examples are provided. Bazzani et al. [2012] proposed a novel descriptor for person reidentification 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 embeds the essence of the textural, shape, and appearance properties of the data [Jojic et al. 2003]. A completely different approach has been adopted by Satta et al. [2012a]. Each individual is represented as a vector of dissimilarity values from a set of learned visual prototypes. Even if the reidentification accuracy is lower than other approaches, particularly when the number of prototypes is low, the tradeoff between processing time and accuracy is still advantageous. This presents us with an application for real-time scenarios.

Distance metric: In the past, machine learning has been frequently neglected by reidentification as the majority of reviewed papers seemingly apply common distance metrics and nearest-neighbor approaches to reidentify the same person (e.g., [Gandhi and Trivedi 2006; Bak et al. 2010; Baltieri et al. 2010]). 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., [Farenzena et al. 2010].) Recently, more attention has been devoted to learning a good metric. Dikmen et al. [2011] proposed an SVM framework to obtain an optimized metric for nearest-neighbor classification called Large Margin Nearest Neighbor (LMNN). Zheng et al. [2011] introduced a novel Probabilistic Relative Distance Comparison (PRDC) model, which differs from most existing distance learning methods in that, rather than minimizing intraclass variation while maximizing interclass variation, it aims to maximize the probability of a pair of true matches having smaller distance than that of a wrong-match pair. An extension of the original method has been presented by the same author in Zheng et al. [2012]. Kuo et al. [2010] and Li et al. [2009b] designed boost learning frameworks to generate an affinity model that is exploited for people’s tracklet association. Similarly, Yang et al. [2011] handled the association problem using a CRF model. In Hirzer et al. [2012], 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 similar to the Mahalanobis distance, providing a dissimilarity score between two feature vectors.

3.5. Spatial Level Mapping on a Body Model

People reidentification 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.

2D body model: 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* [Lantagne et al. 2003; Farenzena et al. 2010]). 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 [Orwell et al. 1999]. These holistic features have the advantages of all aggregated measures: reduced sensitivity to noise, low computational cost, and no alignment or segmentation steps 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 that 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 [Liu and Yang 2009]).

Among others, the *cylindrical* shape and the *legs-torso-head* structure are the most widely utilized body models in surveillance and forensics. By modeling a person as a cylindrical shape (or, more generally, as a solid of revolution), the horizontal variations of the 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. [2005] divided the person's silhouette into 10 horizontal stripes with the mean color of each stripe being stored as a 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 [Albu et al. 2006; Monari et al. 2009; Park et al. 2006; Lantagne et al. 2003; Gong et al. 2011]. Albu et al. [2006] placed the cuts at 30% and 80% of the total height, while Monari et al. [2009] opted to make cuts at the 15% and 70% mark, respectively. Other methods did not divide the RoI into fixed parts but propose alternative solutions. Farenzena et al. [2010] automatically computed the cut points from profile histograms and split the torso and legs into two parts using a symmetry-based algorithm. Cheng et al. [2011] adopted the pictorial structure technique proposed by Andriluka et al. [2009] 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. [2010] 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 provides an illustration of the examples mentioned earlier. The body models are also reported.

3D body model: Over the past few years, 3D models are drawing an ever increasing interest by the field of people surveillance, particularly with regard to the reidentification task. The first attempt to incorporate a 3D body model was done by Gandhi and Trivedi [2006], 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 [Andriluka et al. 2010; Colombo et al. 2008b]. 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. Recently, a new monolithic 3D model called SARC3D [Baltieri et al. 2011c] has been proposed. This model leaves us with a more uninvolved procedure where data can be processed in real time with embedded color appearance features (see Figure 5(b)). The model aims to store local features such as color histograms and texture descriptors together with their spatial locations on a 3D body model.

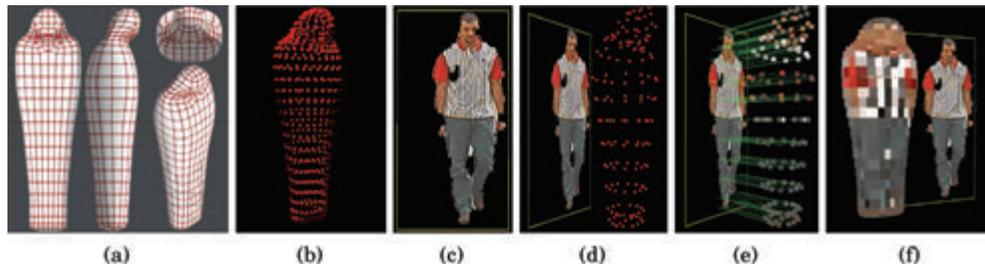


Fig. 5. Sarc3D model [Baltieri et al. 2011c]: (a) lateral, top, and frontal view of the 3D body hull; (b) the vertex set generated sampling the 3D hull; (c–f) steps for the computation of the 3D signature from a 2D image: (c) the silhouette is provided as RoI, (d) the silhouette is aligned to 3D body model, (e) the color features are locally mapped to the body model, and (f) after the integration of multiple shots the final model is reached.

3.6. 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 reidentification 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 intracamera object tracking. Data is collected and merged in a “late-fusion”-like manner rather than being integrated before the tracking.

Consistent labeling approaches proposed by Khan and Shah [2003], Calderara et al. [2008a], and, more recently, Lian et al. [2010] and Madrigal and Hayet [2011] belong to this category, as all methods require at least a partial overlapping among the camera fields of view. Baltieri et al. [2011a] adopted a multicamera system to track people in the overlapping area being surveyed, while the 3D model-based reidentification proposed by Baltieri et al. [2011c] joins together the track segments corresponding to the same person entering once more in the central zone.

Image retrieval: In forensics, the real-time constraint is no longer problematic since the computation is bordering being offline and human interaction is permissible. Given a query item, all the frames/images corresponding to the same person should be retrieved. The reidentification 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 Bird et al. [2005] Metternich et al. [2010] can be applied to any resolution. The method of Farenzena et al. [2010] requires medium resolution to find symmetry axes and texture patterns, while the work of Fischer et al. [2011] 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 [Andriluka et al. 2008] tries to link the provided detections of the same person by means of reidentification algorithms. In this case, the reidentification task works in terms of *short-term tracking*, similar to a “pure” reidentification approach



Fig. 6. Shot examples from (a) ViPER [Gray et al. 2007] and (b) ETHZ for reidentification [Schwartz and Davis 2009]. ViPER contains a couple of cropped images for each person, while ETHZ is composed by full frames from the original dataset [<http://www.vision.ee.ethz.ch/~aess/dataset>] (left) and the bounding box annotation to crop the person images (right).

that requires a feature-based signature for each detection. Breitenstein et al. [2009] compared a set of different color (RGI/RGB/HS/Lab) and texture (LBP/Haar) features, while Brendel et al. [2011] 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 [Perera et al. 2006], network flow [Zhang et al. 2008], or spectral clustering [Coppi et al. 2011]. However, the view or temporal gap assumed by the definition of reidentification (see Section 1) should be null or at least limited. This application scenario is very close to tracking, and so the following section and Table II illustrate some specific examples. We also refer the reader to specific surveys on the topic, such as the work by Yilmaz et al. [2006].

4. DATASETS AND METRICS FOR PERFORMANCE EVALUATION

4.1. Datasets

While several datasets are publicly available for testing camera tracking, action classification systems, or surveillance (see a short review by Vezzani and Cucchiara [2010]) and multimedia [Over et al. 2011], few can actually be adopted for reidentification evaluation, especially for multiple-shot techniques or 3D body models.

ViPER: Currently, one of the most popular and challenging datasets to test people reidentification as image retrieval is ViPER [Gray et al. 2007], which contains 632 pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions (see Figure 6(a) <http://www.vision.ee.ethz.ch/~aess/dataset>). The dataset was collected in an academic setting over the course of several months and each image is scaled to 128×48 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 Figure 6(a) [<http://www.vision.ee.ethz.ch/~aess/dataset>]. Currently, the best results on this dataset have been obtained by Farenzena et al. [2010] on a subset of the dataset and Gray and Tao [2008], who are the dataset's original authors. ViPER cannot 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 the following section.

I-LIDS: The I-LIDS Multiple-Camera Tracking Scenario (MCTS) [Nilski 2008] was captured inside a busy airport arrival hall. With an average of four images for each person, it contains a total of 476 shots of 119 people captured by multiple nonoverlapping 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. [2011], Prosser et al. [2010], and Zheng et al. [2011] for a performance evaluation of their proposal.

CAVIAR4REID: This is a small dataset specifically designed for evaluating person reidentification algorithms by Cheng et al. [2011]. 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 of each pedestrian. For each pedestrian, a set of images by 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 reidentification.

ETHZ: The ETHZ dataset for appearance-based modeling was generated by Schwartz and Davis [2009] from the original ETHZ video dataset [Ess et al. 2007]. The original ETHZ dataset was used for human detection and is composed of four video sequences. Samples of testing sequence frames are shown in Figure 6(b) [<http://www.vision.ee.ethz.ch/~aess/dataset>]. The ETHZ dataset presents the additional challenge being captured by moving cameras. This camera setup provides a range of variations in people's appearances, with significant changes in pose and illumination.

ViSOR dataset: This dataset was introduced as a means of testing multiple-shot 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 correspond to predefined positions and postures of the people that were manually selected. The annotated dataset 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 Figure 7).

3DPeS dataset: Recently, the 3DPeS dataset was published [Baltieri et al. 2011b] with the aim of overcoming limitations from the datasets previously mentioned. 3DPeS provides a large volume of data that tests all the usual steps in video surveillance, such as segmentation and tracking.

The dataset is captured by a real surveillance setup and is composed of eight different surveillance cameras (Figure 8) monitoring an area of the University of Modena and Reggio Emilia's (UNIMORE) campus. Data was collected over the course of several days. 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.

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 \times 2 hours/day \times 5 cameras). Approximately 44 individuals were detected and matched through the five 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 multicamera system and contains sequences with different crowd activities in a



Fig. 7. Sample silhouettes from ViSOR reidentification dataset.

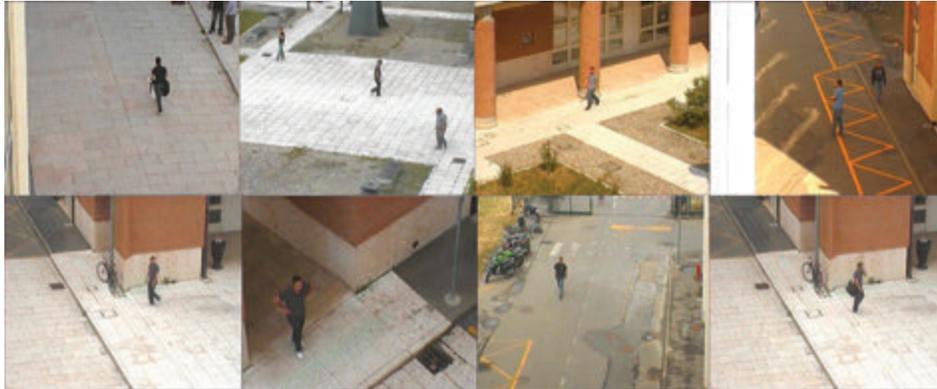


Fig. 8. Sample frames from 3DPeS.

real-world environment. Each sequence involves a subset of eight available cameras and up to approximately 40 actors.

Videoweb Activities Dataset [Denina et al. 2011]: The Videoweb Activities Dataset is composed of roughly 2.5 hours of video footage taken by multicamera systems in realistic scenarios and contains people performing numerous repetitive and nonrepetitive tasks. Data was collected over 4 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 reidentification purposes, this dataset presents us with six synchronized views acquired by six full-HD cameras during a soccer match [D’Orazio et al. 2009]. The high similarity among players of the same team makes the intraview tracking and the reidentification 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 II. For additional references to surveillance datasets, please refer to Vezzani and Cucchiara [2010] or the Cantata Project repository [project 2008].

The main benchmarks for reidentification are summarized in Table III.

4.2. 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 [Frontex 2011], different metrics are available that relate the specific implementation of reidentification as identification or recognition.

Reidentification 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. Different from content-based retrieval problems, where there are relatively few clusters and very large amounts 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 [Amigó et al. 2009]. *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 reidentified.

The performance evaluation of reidentification 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 [Hamdoun et al. 2008].

Tasks of reidentification in long-term tracking also fall into this category, especially in surveillance with a network of overlapped or disjoint cameras. With a tracking system, the reidentification algorithm should generate tracks for as long as possible, avoiding errors such as identity switch, erroneous split and merge of tracks, and over- and undersegmentation of traces. For detection and tracking purposes, the ETISEO project [Nghiem et al. 2007] proposed some metrics that could potentially be adopted in reidentification. 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 multicamera tracking algorithm, but a problem exists where the evaluation results depend not only on the reidentification algorithms but also on the detection and single-camera tracking. The second metric regards the reidentification precision, evaluating how many identities have been assigned to the same real person.

Finally, let us cite the work of Leung et al. [2008] about performance evaluation of reacquisition 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 reidentification system can reduce the entropy of the surveillance framework.

Reidentification as recognition: In this category, the reidentification task aims to provide a set of ranked items given a query target, with the main hypothesis being that 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 reidentification process could

Table III. Datasets Available for People Reidentification

Name & Ref	Image/Video	People	Additional info
ViPER [Gray et al. 2007]	Still Images	632	Scenario: Outdoor Place: Outdoor surveillance People Size: 128×48 vision.soe.ucsc.edu
I-LIDS [Nilski 2008]	Video [fps=25] 5 cameras PAL	1000	Scenario: Outdoor/Indoor Place: Collection from different scenarios People Size: 21×53 to 176×326 www.ilids.co.uk
I-LIDS-MA [Nilski 2008]	Still Images PAL	40	Scenario: Indoor Place: Airport People Size: 21×53 to 176×326 www.ilids.co.uk
I-LIDS-AA [Nilski 2008]	Still Images PAL	119	Scenario: Indoor Place: Airport People Size: 21×53 to 176×326 www.ilids.co.uk
CAVIAR4REID [Cheng et al. 2011]	Still Images 384×288	72	Scenario: Indoor Place: Shopping centre People Size: 17×39 to 72×144 www.lorisbazzani.info
ETHZ [Schwartz and Davis 2009]	Video [fps=15] 1 cameras 640×480	146	Scenario: Outdoor Place: Moving cameras on city street People Size: 13×30 to 158×432 http://homepages.dcc.ufmg.br/~william/
ViSOR dataset [Baltieri et al. 2011c]	Still Images 704×576	50	Scenario: Outdoor Place: University Campus People Size: 54×187 to 149×306 www.openvisor.org
3DPeS dataset [Baltieri et al. 2011b]	Video [fps=15] 8 cameras 704×576	200	Scenario: Outdoor Place: University Campus People Size: 31×100 to 176×267 www.openvisor.org
TRECVID 2008 [Smeaton et al. 2006]	Video [fps=25] 5 cameras PAL	300	Scenario: Indoor Place: Gatwick International Airport - London People Size: 21×53 to 176×326 www-nlpir.nist.gov/projects/tv2008/
PETS2009 [PETS 2009]	Video [fps=7] 8 cameras 768×576	40	Scenario: Outdoor Place: Outdoor surveillance People Size: 26×67 to 57×112 www.cvg.rdg.ac.uk/PETS2009/
Videoweb Activities Dataset [Denina et al. 2011]	Video [fps=30] 8 cameras 640×480	16	Scenario: Outdoor Place: Courtyard and intersections People Size: 32×62 to 86×170 www.ee.ucr.edu/~amitr
ISSIA Soccer dataset [D’Orazio et al. 2009]	Video [fps=25] 6 cameras 1920×1080	25	Scenario: Outdoor Place: Soccer match People Size: 42×82 to 57×130 www.issia.cnr.it/soccerdataset.html

Table IV. Quantitative Comparison of Some Methods on the ViPER Dataset
 The reported values have been sampled from the CMC curves at ranks 1, 2, 5, and 20;
 see Section 4.2.

Method	Rank-1	Rank-5	Rank-10	Rank-20
RGB Histogram	0,04	0,11	0,20	0,27
ELF [Gray and Tao 2008]	0,08	0,24	0,36	0,52
Shape and Color Covariance Matrix [Metternich et al. 2010]	0,11	0,32	0,48	0,70
Color-SIFT [Metternich et al. 2010]	0,05	0,18	0,32	0,52
SDALF [Farenzena et al. 2010]	0,20	0,39	0,49	0,65
Ensemble-RankSVM [Prosser et al. 2010]	0,16	0,38	0,53	0,69
PRDC [Zheng et al. 2011]	0,15	0,38	0,53	0,70
MCC [Zheng et al. 2011]	0,15	0,41	0,57	0,73
IML [Ali et al. 2010]				
- 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

be considered a ranking problem [Gray et al. 2007] where the *Cumulative Matching Characteristic* (CMC) curve is the proper performance evaluation metric [Moon and Phillips 2001], 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. Given a set of M query samples q_i , let $\mathbf{T} = (t_1 \dots t_N)$ be the list of the N test images ordered by the reidentification score. Let r_i be the index of the image within \mathbf{T} that matches with q_i . Thus, the CMC curve is obtained by plotting the following values:

$$CMC(n)_{n=1\dots N} = \frac{1}{M} \cdot \sum_{i=1\dots M} \begin{cases} 1 & r_i \leq n \\ 0 & r_i > n \end{cases} \quad (1)$$

To demonstrate the use of the CMC curve, we provide a comparison of two state-of-the-art solutions tested on the 3DPes dataset. In particular, we compared the SDALF algorithm by Farenzena et al. [2010] (silhouette as RoI, single/multiple shots, color, shape, and texture features mapped on a 2D symmetry-based body model) against Sarc3D by Baltieri et al. [2011c] (silhouette as RoI, multiple shots, color features mapped on a 3D body model). To ensure a fair comparison was made, appearance images were automatically extracted at random from each video sequence and analyzed by the reidentification algorithms. In the case of single-shot SDALF (SDALF-SS), we randomly selected a single-view shot from each video sequence. Between three and five images for each sequence were selected for the multiple-shot implementations (SDALF-MS and Sarc3D).

Ten test runs from sequences of 56 randomly selected people from the dataset are performed on each system. Table IV illustrates the accuracy of three methods (among others), while Figure 9 shows some selected test queries. As shown in the table, the use of multiple shots and 3D body models enhances reidentification performance. While the SDALF approach produces slightly less consistent results, its ease of applicability even when no information about camera setting is given is what makes this such a pragmatic model. Both methods are very time consuming and the Sarc3D method also requires a specific 3D alignment step.

Since the adopted definition of reidentification as recognition given by Frontex [2011] 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) [Jungling and

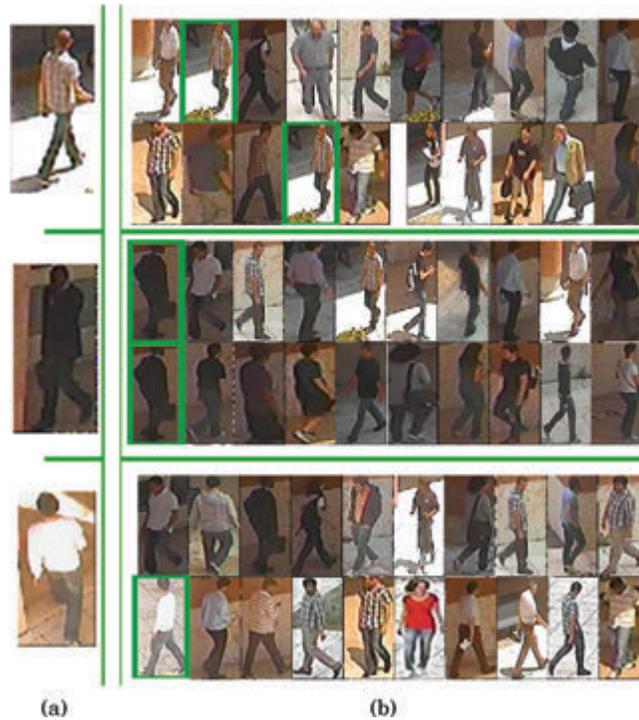


Fig. 9. Three example queries using the 3DPes dataset. (a) Probe image (for Sarc3D this is just one of the images used for the model creation). (b) The top 10 results (sorted left to right) using SDALF (top) and Sarc3D results. The correct match is highlighted in green.

Arens 2010], and the Decision-Error Tradeoff (DET) curve can be evaluated, whether or not two snapshots are associated to the same person.

Reidentification 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 [Meuwly 2006]. With this aim, a Likelihood Ratio (LR) has been suggested for solving different forensics problems like speaker identification [Gonzalez-rodriguez et al. 2003], DNA analysis [Balding 2005], and face recognition [Ali et al. 2010].

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.

$$LR = \frac{P(A|B)}{P(A|\neg B)} \quad (2)$$

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

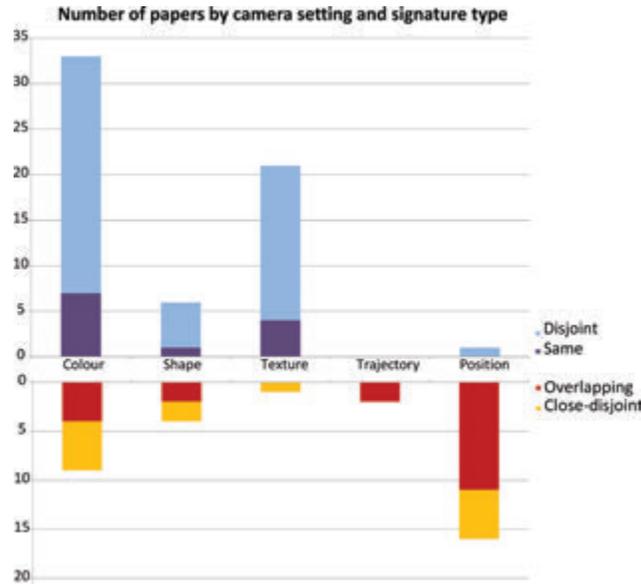


Fig. 10. Histogram of the features used in the reviewed approaches.

evidence if an unidentified person is supposed to be the source. Similar discussions were introduced in a survey for face recognition in forensics [Ali et al. 2010].

5. DISCUSSION

In this section, our goal is to analyze the direction research has tended toward in the last few years. As highlighted in Figure 10, many proposed methods share common templates. First, the camera setting affects the choice of signature and the methods of similarity assessment. The geometric position is most reliable if the camera's fields of view are overlapped or partially overlapped. 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 [2005]. 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 [Madden et al. 2007; Prosser et al. 2008].

Only recently have more complex signatures based on texture been proposed and with very promising results, more so with SIFT [Teixeira and Corte-Real 2009; Bak et al. 2010] or SURF [Hamdoun et al. 2008] descriptors (see Figure 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 models have been proposed by Farenzena et al. [2010], Doretto et al. [2011], and Gandhi and Trivedi [2006]. Farenzena et al. [2010] take 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

Table V. Quantitative Comparison of Some Methods Using Frames from CAVIAR
The reported values have been sampled from the CMC curves at ranks 1, 2, 5, and 20;
see Section 4.2.

Method	Rank-1	Rank-5	Rank-10	Rank-20
<i>Caviar4Reid dataset</i>				
AHPE [Bazzani et al. 2012]	0,08	0,32	0,53	0,74
SDALF [Farenzena et al. 2010]	0,09	0,38	0,58	0,77
CPS [Cheng et al. 2011]	0,16	0,48	0,69	0,86
<i>Personal dataset</i>				
BoF+SVM [Liu and Yang 2009]	0,58	0,93	0,98	1,00

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. [2011] 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 [2006] and the recently proposed SARC3D body model [Baltieri et al. 2011c] 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 Chen et al. [2011]) should be very useful to improve both accuracy and speed.

Table IV and Table V contain quantitative comparisons of some of the reviewed techniques 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. By excluding the last three rows obtained with interactive refinements through relevance feedback, the best Rank-1 performance on the ViPER dataset are provided by Farenzena et al. [2010]. Sophisticated ranking strategies could improve the performance of Rank-5, Rank-10, and Rank-20, as highlighted in the results reported by Prosser et al. [2010] and Zheng et al. [2011]. 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.

Over the last few years, the role of machine learning has become a central component in the computer vision field. People reidentification is aligned with this trend and current research is primarily devoted to the learning aspects of this area. As described in Section 3.4, specific feature sets like SDALF [Farenzena et al. 2010] have been replaced by a set of standard color and texture descriptors [Prosser et al. 2010; Ali et al. 2010; Zheng et al. 2011; Liu et al. 2012; Zheng et al. 2012; Mazzon et al. 2012]. The selection of important features is then processed by a machine learning algorithm in an explicit [Liu et al. 2012] or implicit way [Li et al. 2012; Zheng et al. 2012].

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 reidentification. Limitations mentioned previously that past research has failed to overcome can now be eradicated. Articulated motion capture systems (addressed in the past on color images [Deutscher and Reid 2005; Salzmann and Urtasun 2010; Pons-Moll et al. 2011] and, more recently, on depth streams [Shotton et al. 2011; Taylor et al. 2012]) are nowadays feasible in real time. In exploiting this information, new methods for reidentification 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. [2012] defined a set of ratios of joint distances as a person signature. Albiol et al. [2012] took advantage of the body model to generate a color histogram for each vertical stripe. Different from the methods presented in the past [Bird et al. 2005] and described in Section 3.5, 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 does not allow yet acceptable performance. These methods still require extensive research.

Finally, the recent work of Layne et al. [2012] proposing an attribute-based approach is worthy of mention. Taking inspiration from the operating procedures of human experts [Vaquero et al. 2009], they moved the reidentification 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 also can be used when a description is provided as a verbal identikit.

6. CONCLUSIONS

In this article, we focused on the new problem of people reidentification. We addressed the problem by proposing a multidimensional categorization of the proposed approaches, the related design, and computational aspects and the resulting challenges. We proposed a classification based on the application scenario, aspects of camera setting, sample set cardinality, and region of interest. We also included signature type together with its possible local or global mapping to a human body model. We accordingly reviewed more than 100 papers in a survey. Finally, we presented benchmark datasets and discussed possible evaluation metrics.

REFERENCES

- AGGARWAL, J. K. AND CAI, Q. 1999. Human motion analysis: A review. *Comput. Vis. Image Understanding* 73, 3, 428–440.
- ALAH, A., VANDEREGHEYNST, P., BIERLAIRE, M., AND KUNT, M. 2010. Cascade of descriptors to detect and track objects across any network of cameras. *Comput. Vis. Image Understanding* 114, 6, 624–640.
- ALBIOL, A., ALBIOL, A., OLIVER, J., AND MOSSI, J. 2012. Who is who at different cameras: People re-identification using depth cameras. *IET Comput. Vision* 6, 5, 378–387.
- ALBU, A., LAURENDEAU, D., COMTOIS, S., OUELLET, D., HEBERT, P., ZACCARIN, A., PARIZEAU, M., BERGEVIN, R., MALDAGUE, X., DROUIN, R., DROUIN, S., MARTEL-BRISSE, N., JEAN, F., TORRESAN, H., GAGNON, L., AND LALIBERTE, F. 2006. MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras. In *Proceedings of the International Conference on Pattern Recognition*. IEEE, 924–928.
- ALI, S., JAVED, O., HAERING, N., AND KANADE, T. 2010. Interactive retrieval of targets for wide area surveillance. In *Proceedings of the ACM International Conference on Multimedia (MM'10)*. ACM, New York, 895–898.
- ALI, T., VELDTHUIS, R., AND SPREEUWERS, L. 2010. Forensic face recognition: A survey. Tech. rep. TRC-CTIT-10-40. Centre for Telematics and Information Technology, University of Twente, Enschede.
- AMIGÓ, E., GONZALO, J., ARTILES, J., AND VERDEJO, F. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retrieval* 12, 461–486.
- ANDRILUKA, M., ROTH, S., AND SCHIELE, B. 2008. People-tracking-by-detection and people-detection-by-tracking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1–8.
- ANDRILUKA, M., ROTH, S., AND SCHIELE, B. 2009. Pictorial structures revisited: People detection and articulated pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1014–1021.
- ANDRILUKA, M., ROTH, S., AND SCHIELE, B. 2010. Monocular 3d pose estimation and tracking by detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 623–630.
- ANJUM, N. AND CAVALLARO, A. 2009. Trajectory association and fusion across partially overlapping cameras. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 201–206.
- AZZ, K.-E., MERAD, D., AND FERTIL, B. 2011. People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 303–308.
- BABENKO, B., YANG, M.-H., AND BELONGIE, S. 2009. Visual tracking with online multiple instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 983–990.

- BAK, S., CORVEE, E., BREMOND, F., AND THONNAT, M. 2010. Person re-identification using spatial covariance regions of human body parts. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 435–440.
- BAK, S., CORVEE, E., BREMOND, F., AND THONNAT, M. 2011. Multiple-shot human re-identification by mean riemannian covariance grid. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 179–184.
- BALDING, D. 2005. *Weight-of-Evidence for Forensic DNA Profiles*. Wiley.
- BALTIERI, D., UTASI, A., VEZZANI, R., CSABA, B., SZIRANYI, T., AND CUCCHIARA, R. 2011a. Multi-view people surveillance using 3D information. In *Proceedings of the 11th International Workshop on Visual Surveillance 2011*. 1817–1824.
- BALTIERI, D., VEZZANI, R., AND CUCCHIARA, R. 2010. 3D body model construction and matching for real time people re-identification. In *Proceedings of the Eurographics Italian Chapter Conference 2010 (EG-IT'10)*.
- BALTIERI, D., VEZZANI, R., AND CUCCHIARA, R. 2011b. 3DPeS: 3D people dataset for surveillance and forensics. In *Proceedings of the 1st International ACM Workshop on Multimedia Access to 3D Human Objects*.
- BALTIERI, D., VEZZANI, R., AND CUCCHIARA, R. 2011c. Sarc3d: A new 3d body model for people tracking and re-identification. In *Proceedings of the IEEE International Conference on Image Analysis and Process*. 197–206.
- BARBOSA, I. B., CRISTANI, M., BUE, A. D., BAZZANI, L., AND MURINO, V. 2012. Re-identification with rgb-d sensors. In *Proceedings of the 1st International ECCV Workshop on Re-Identification (ReID'12)*, A. Fusiello, V. Murino, and R. Cucchiara, Eds., Lecture Notes in Computer Science Series, vol. 7583, Springer, 433–442.
- BAUML, M. AND STIEFELHAGEN, R. 2011. Evaluation of local features for person re-identification in image sequences. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 291–296.
- BAZZANI, L., CRISTANI, M., PERINA, A., AND MURINO, V. 2012. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognit. Letters*. 33, 7, 898–903.
- BERCLAZ, J., FLEURET, F., AND FUA, P. 2006. Robust people tracking with global trajectory optimization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. Vol. 1. 744–750.
- BIRCHFIELD, S. AND RANGARAJAN, S. 2005. Spatiograms versus histograms for region-based tracking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. Vol. 2. 1158–1163.
- BIRD, N., MASOUD, O., PAPANIKOLOPOULOS, N., AND ISAACS, A. 2005. Detection of loitering individuals in public transportation areas. *IEEE Trans. Intell. Transp. Syst.* 6, 2, 167–177.
- BLACK, J., ELLIS, T., AND MAKRIKIS, D. 2004. Wide area surveillance with a multi camera network. *IEE Semin. Digests 2004*, 10426, 21–25.
- BLACK, J., ELLIS, T., AND ROSIN, P. 2002. Multi view image surveillance and tracking. In *Proceedings of the Workshop on Motion and Video Computing, 2002. IEEE Comput. Soc*, 169–174.
- BOWDEN, R. AND KAEWTRAKULPONG, P. 2005. Towards automated wide area visual surveillance: Tracking objects between spatially-separated, uncalibrated views. *IEE Proc. Vision, Image Signal Process.* 152, 2, 213–223.
- BREITENSTEIN, M., REICHLIN, F., LEIBE, B., KOLLER-MEIER, E., AND VAN GOOL, L. 2009. Robust tracking-by-detection using a detector confidence particle filter. In *Proceedings of the IEEE International Conference on Computer Vision*. 1515–1522.
- BRENDEL, W., AMER, M., AND TODOROVIC, S. 2011. Multiobject tracking as maximum weight independent set. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1273–1280.
- CAI, Q. AND AGGARWAL, J. 1998. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In *Proceedings of the IEEE International Conference on Computer Vision*. 356–362.
- CAI, Q. AND AGGARWAL, J. 1999. Tracking human motion in structured environments using a distributed-camera system. *IEEE Trans. Pattern Anal. Mach. Intell.* 21, 11, 1241–1247.
- CAI, Q. AND AGGARWAL, J. K. 1996. Tracking human motion using multiple cameras. In *Proceedings of the International Conference on Pattern Recognition*. Vol. 3. IEEE Computer Society, Los Alamitos, CA, 68.
- CALDERARA, S., CUCCHIARA, R., AND PRATI, A. 2008a. Bayesian-competitive consistent labeling for people surveillance. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 2, 354–360.
- CALDERARA, S., PRATI, A., AND CUCCHIARA, R. 2008b. HECOL: Homography and epipolar-based consistent labeling for outdoor park surveillance. *Comput. Vis. Image Understanding* 111, 1, 21–42.

- CHANG, T.-H. AND GONG, S. 2001. Tracking multiple people with a multi-camera system. In *Proceedings of the IEEE Workshop Multi-Object Tracking*. IEEE, 19–26.
- CHEN, C., HEILLI, A., AND ODOBEZ, J. 2011. Combined estimation of location and body pose in surveillance video. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 5–10.
- CHEN, K.-W., LAI, C.-C., HUNG, Y.-P., AND CHEN, C.-S. 2008. An adaptive learning method for target tracking across multiple cameras. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1–8.
- CHENG, D. S., CRISTANI, M., STOPPA, M., BAZZANI, L., AND MURINO, V. 2011. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC'11)*.
- COLOMBO, A., ORWELL, J., AND VELASTIN, S. 2008a. Colour constancy techniques for re-recognition of pedestrians from multiple surveillance cameras. In *Proceedings of the Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2'08)*.
- COLOMBO, C., DEL BIMBO, A., AND VALLI, A. 2008b. A real-time full body tracking and humanoid animation system. *Parallel Comput.* 34, 718–726.
- CONG, D. N. T., KHOUDOUR, L., AND ACHARD, C. 2010a. People reacquisition across multiple cameras with disjoint views. In *Proceedings of the International Conference on Image and Signal Processing (ICISP'10)*. Springer-Verlag, Berlin, 488–495.
- CONG, D. N. T., KHOUDOUR, L., ACHARD, C., MEURIE, C., AND LEZORAY, O. 2010b. People re-identification by spectral classification of silhouettes. *Signal Process.* 90, 8, 2362–2374.
- CONTE, D., FOGGIA, P., PERCANNELLA, G., AND VENTO, M. 2011. A multiview appearance model for people re-identification. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 297–302.
- COPPI, D., CALDERARA, S., AND CUCCHIARA, R. 2011. Appearance tracking by transduction in surveillance scenarios. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*.
- DALAL, N., TRIGGS, B., AND SCHMID, C. 2006. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference Computer Vision*. Springer.
- DANTCHEVA, A. AND DUGELAY, J.-L. 2011. Frontal-to-side face re-identification based on hair, skin and clothes patches. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 309–313.
- DANTCHEVA, A., DUGELAY, J.-L., AND ELIA, P. 2010. Soft biometrics systems: Reliability and asymptotic bounds. In *Proceedings of the IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS'10)*. 1–6.
- DANTCHEVA, A., VELARDO, C., D'ANGELO, A., AND DUGELAY, J.-L. 2011. Bag of soft biometrics for person identification—new trends and challenges. *Multimedia Tools Appl.* 51, 2, 739–777.
- DE OLIVEIRA, I. O. AND PIO, J. L. S. 2009. People reidentification in a camera network. In *Proceedings of the 2nd International Conference on Computer Science and Its Applications*. IEEE, 1–8.
- DELAC, K. AND GRGIC, M. 2004. A survey of biometric recognition methods. In *Proceedings of the International Symposium Electronics in Marine (ELMAR'04)*. 184–193.
- DENINA, G., BHANU, B., NGUYEN, H. T., DING, C., KAMAL, A., RAVISHANKAR, C., ROY-CHOWDHURY, A., IVERS, A., AND VARDA, B. 2011. *VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication*. Springer, London, 335–347.
- DENMAN, S., FOOKES, C., BIALKOWSKI, A., AND SRIDHARAN, S. 2009. Soft-biometrics: Unconstrained authentication in a surveillance environment. In *Proceedings of the 2009 Digital Image Computing: Techniques and Applications (DICTA'09)*. IEEE Computer Society, Washington, DC, 196–203.
- DEUTSCHER, J. AND REID, I. 2005. Articulated body motion capture by stochastic search. *Int. J. Comput. Vision* 61, 2, 185–205.
- DIKMEN, M., AKBAS, E., HUANG, T. S., AND AHUJA, N. 2011. Pedestrian recognition with a learned metric. In *Proceedings of the 10th Asian Conference on Computer Vision (ACCV'10), Part IV*. Springer-Verlag, Berlin, 501–512.
- D'ORAZIO, T., LEO, M., MOSCA, N., SPAGNOLO, P., AND MAZZEO, P. 2009. A semi-automatic system for ground truth generation of soccer video sequences. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 559–564.
- DOROTTO, G., SEBASTIAN, T., TU, P. H., AND RITTSCHER, J. 2011. Appearance-based person reidentification in camera networks: Problem overview and current approaches. *J. Ambient Intell. Humanized Comput.* 2, 2, 127–151.
- DUTAGACI, H., SANKUR, B., AND YÖRÜK, E. 2008. Comparative analysis of global hand appearance-based person recognition. *J. Electron. Imaging* 17, 1, 1–19.

- ELLIS, T. AND BLACK, J. 2003. A multi-view surveillance system. In *Proceedings of the IEEE Symposium on Intelligence Distributed Surveillance Systems*. 11/1–11/5.
- ESS, A., LEIBE, B., AND GOOL, L. V. 2007. Depth and appearance for mobile scene analysis. In *Proceedings of the IEEE International Conference on Computer Vision*.
- FARENZENA, M., BAZZANI, L., PERINA, A., MURINO, V., AND CRISTANI, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 2360–2367.
- FISCHER, M., EKENEL, H., AND STIEFELHAGEN, R. 2011. Person re-identification in tv series using robust face recognition and user feedback. *Multimedia Tools Appl.* 55, 83–104. 1
- FORSYTH, D. A. AND PONCE, J. 2002. *Computer Vision: A Modern Approach* 1st Ed. Prentice Hall.
- FRONTEX. 2011. Application of surveillance tools to border surveillance—Concept of operations. http://ec.europa.eu/enterprise/policies/security/files/doc/conops_gmes_en.pdf.
- GANDHI, T. AND TRIVEDI, M. 2006. Panoramic appearance map (PAM) for multi-camera based person re-identification. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance*. IEEE, 78–78.
- GHEISSARI, N., SEBASTIAN, T. B., AND HARTLEY, R. 2006. Person reidentification using spatiotemporal appearance. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. Vol. 2. 1528–1535.
- GIJSENIJ, A., GEVERS, T., AND VAN DE WEIJER, J. 2011. Computational color constancy: Survey and experiments. *IEEE Trans. Image Process.* 20, 9, 2475–2489.
- GILBERT, A. AND BOWDEN, R. 2006. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *Proceedings of the European Conference on Computer Vision*. 125–136.
- GONG, H., SIM, J., LIKHACHEV, M., AND SHI, J. 2011. Multi-hypothesis motion planning for visual object tracking. In *Proceedings of the IEEE International Conference Computer Vision*. 619–626.
- GONZALEZ-RODRIGUEZ, J., FIERREZ-AGUILAR, J., AND ORTEGA-GARCIA, J. 2003. Forensic identification reporting using automatic speaker recognition systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*. 93–96.
- GORELICK, L., BLANK, M., SHECHTMAN, E., IRANI, M., AND BASRI, R. 2007. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 12, 2247–2253.
- GRABNER, H., MATAS, J., VAN GOOL, L., AND CATTIN, P. 2010. Tracking the invisible: Learning where the object might be. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1285–1292.
- GRAY, D., BRENNAN, S., AND TAO, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS'07)*.
- GRAY, D. AND TAO, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference Computer Vision*. 262.
- GUALDI, G., PRATI, A., AND CUCCHIARA, R. 2011. A multi-stage pedestrian detection using monolithic classifiers. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*.
- HAMDOUN, O., MOUTARDE, F., STANCIULESCU, B., AND STEUX, B. 2008. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proceedings of the International Conference on Distributed Smart Cameras*. IEEE, 1–6.
- HARTLEY, R. I. AND ZISSERMAN, A. 2004. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press.
- HAVASI, L., SZLAVIK, Z., AND SZIRANYI, T. 2005. Eigenwalks: Walk detection and biometrics from symmetry patterns. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, III–289.
- HIRZER, M., ROTH, P. M., KÖSTINGER, M., AND BISCHOF, H. 2012. Relaxed pairwise learned metric for person re-identification. In *Proceedings of the Conference on Computer Vision (ECCV'12)*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., Lecture Notes in Computer Science Series, vol. 7577. Springer, Berlin, 780–793.
- HU, L., JIANG, S., HUANG, Q., AND GAO, W. 2008. People re-detection using Adaboost with sift and color correlogram. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 1348–1351.
- HU, W., HU, M., ZHOU, X., TAN, T., LOU, J., AND MAYBANK, S. 2006. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 4, 663–671.
- HUANG, T. AND RUSSELL, S. 1998. Object identification: A Bayesian analysis with application to traffic surveillance. *Artificial Intell.* 103, 1–17.

- HYODO, Y., YUASA, S., FUJIMURA, K., NAITO, T., AND KAMIJO, S. 2008. Pedestrian tracking through camera network for wide area surveillance. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 656–661.
- JAIN, A. K., DASS, S. C., NANDAKUMAR, K., AND N, K. 2004. Soft biometric traits for personal recognition systems. In *Proceedings of the International Conference on Biometric Authentication, Hong Kong*. 731–738.
- JAVED, O. AND SHAFIQUE, K. 2005. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 26–33.
- JAVED, O., SHAFIQUE, K., RASHEED, Z., AND SHAH, M. 2008. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Comput. Vis. Image Understanding* 109, 2, 146–162.
- JING-YING, C., TZU-HENG, W., SHAO-YI, C., AND LIANG-GEE, C. 2008. Spatial-temporal consistent labeling for multi-camera multi-object surveillance systems. In *Proceedings of the IEEE International Symposium on Circuits and Systems*. IEEE, 3530–3533.
- JOJIC, N., FREY, B. J., AND KANNAN, A. 2003. Epitomic analysis of appearance and shape. In *Proceedings of the IEEE International Conference Computer Vision*. 34–43.
- JUNGLING, K. AND ARENS, M. 2010. Local feature based person reidentification in infrared image sequences. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 448–455.
- JUNGLING, K. AND ARENS, M. 2011. View-invariant person re-identification with an implicit shape model. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 197–202.
- KANG, J., COHEN, I., AND MEDIONI, G. 2005. Persistent objects tracking across multiple non overlapping cameras. In *Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05)*. Vol. 2. IEEE, 112–119.
- KETTNAKER, V. AND ZABIH, R. 1999. Bayesian multi-camera surveillance. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 253–259.
- KHAN, S. AND SHAH, M. 2003. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 10, 1355–1360.
- KHAN, S. M. AND SHAH, M. 2009. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 3, 505–519.
- KRUMM, J., HARRIS, S., MEYERS, B., BRUMITT, B., HALE, M., AND SHAFER, S. 2000. Multi-camera multi-person tracking for EasyLiving. In *Proceedings of the 3rd IEEE International Workshop on Visual Surveillance*. IEEE Comput. Soc., 3–10.
- KUO, C.-H., HUANG, C., AND NEVATIA, R. 2010. Multi-target tracking by on-line learned discriminative appearance models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 685–692.
- KUO, C.-H. AND NEVATIA, R. 2011. How does person identity recognition help multi-person tracking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1217–1224.
- LANTAGNE, M., PARIZEAU, M., AND BERGEVIN, R. 2003. VIP: Vision tool for comparing images of people. In *Proceedings of the 16th International Conference on Vision Interface*. 35–42.
- LAYNE, R., HOSPEDALES, T. M., AND GONG, S. 2012. Towards person identification and re-identification with attributes. In *Proceedings of the 1st International ECCV Workshop on Re-Identification (ReID'12)*, A. Fusiello, V. Murino, and R. Cucchiara, Eds., Lecture Notes in Computer Science Series, vol. 7583. Springer, 402–412.
- LEE, L., ROMANO, R., AND STEIN, G. 2000. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 8, 758–767.
- LEUNG, V., ORWELL, J., AND VELASTIN, S. A. 2008. Performance evaluation of re-acquisition methods for public transport surveillance. In *Proceedings of the International Conference on Control, Automation, Robotics and Vision*. IEEE, 705–712.
- LI, Q., CHEN, Q., YU, T., AND LIU, W. 2009a. A P2P camera system with new consistent labeling method involving only simple geometric operations. In *Proceedings of the 11th IEEE International Symposium on Multimedia*. IEEE, 52–56.
- LI, W., WU, Y., MUKUNOKI, M., AND MINOH, M. 2012. Common-near-neighbor analysis for person re-identification. In *Proceedings of the International Conference on Image Processing*. 1621–1624.
- LI, Y., HUANG, C., AND NEVATIA, R. 2009b. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 2953–2960.

- LIAN, G., LAI, J., AND GAO, Y. 2010. People consistent labeling between uncalibrated cameras without planar ground assumption. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 733–736.
- LIN, Z. AND DAVIS, L. S. 2008. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *Proceedings of the 4th International Symposium on Advances in Visual Computing*. 23–34.
- LIU, C., GONG, S., LOY, C. C., AND LIN, X. 2012. Person re-identification: What features are important? In *Proceedings of the 1st International ECCV Workshop on Re-Identification (ReID'12)*, A. Fusiello, V. Murino, and R. Cucchiara, Eds., Lecture Notes in Computer Science Series, vol. 7583. Springer, 391–401.
- LIU, K. AND YANG, J. 2009. Recognition of people reoccurrences using bag-of-features representation and support vector machine. In *Proceedings of the Chinese Conference on Pattern Recognition*. IEEE, 1–5.
- LOKE, Y. R., KUMAR, P., RANGANATH, S., AND HUANG, W. M. 2006. Object matching across multiple non-overlapping fields of view using fuzzy logic. *Acta Automatica Sinica* 36, 6, 978–987.
- MADDEN, C., CHENG, E. D., AND PICCARDI, M. 2007. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Mach. Vision Appl.* 18, 3, 233.
- MADRIGAL, F. AND HAYET, J.-B. 2011. Multiple view, multiple target tracking with principal axis-based data association. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 185–190.
- MAKRIS, D., ELLIS, T., AND BLACK, J. 2004. Bridging the gaps between cameras. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 205–210.
- MAZZON, R., TAHIR, S. F., AND CAVALLARO, A. 2012. Person re-identification in crowd. *Pattern Recognit. Lett.* 33, 14, 1828–1837.
- MEI, X. AND LING, H. 2011. Robust visual tracking and vehicle classification via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 11, 2259–2272.
- METTERNICH, M., WORRING, M., AND SMEULDERS, A. 2010. Color based tracing in real-life surveillance data. *Trans. Data Hiding Multimedia Security V 6010*, 18–33.
- MEUWLY, D. 2006. Forensic individualization from biometric data. *Sci. Justice* 46, 4, 205–213.
- MINDRU, F., TUYTELAARS, T., GOOL, L. V., AND MOONS, T. 2004. Moment invariants for recognition under changing viewpoint and illumination. *Comput. Vis. Image Understanding* 94, 1, 3.
- MIYAZAWA, K., ITO, K., AOKI, T., KOBAYASHI, K., AND NAKAJIMA, H. 2008. An effective approach for iris recognition using phase-based image matching. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 10, 1741–1756.
- MONARI, E., MAERKER, J., AND KROSCHEL, K. 2009. A robust and efficient approach for human tracking in multi-camera systems. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. IEEE, 134–139.
- MOON, H. AND PHILLIPS, P. J. 2001. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception* 30, 303–321.
- NAKAJIMA, C., PONTIL, M., HEISELE, B., AND POGGIO, T. 2003. Full-body person recognition system. *Pattern Recognit.* 36, 9, 1997–2006.
- NGHIEM, A., BREMOND, F., THONNAT, M., AND VALENTIN, V. 2007. Etiseo, performance evaluation for video surveillance systems. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 476–481.
- NILSKI, A. 2008. Evaluating multiple camera tracking systems—The i-lids 5th scenario. In *Proceedings of the 42nd Annual IEEE International Carnahan Conference on Security Technology (ICCST'08)*. 277–279.
- NIU, C. AND GRIMSON, E. 2006. Recovering non-overlapping network topology using far-field vehicle tracking data. In *Proceedings of the International Conference on Pattern Recognition* 4, 944–949.
- OREIFEJ, O., MEHRAN, R., AND SHAH, M. 2010. Human identity recognition in aerial images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 709–716.
- ORWELL, J., REMAGNINO, P., AND JONES, G. 1999. Multi-camera colour tracking. In *Proceedings of the IEEE Workshop on Visual Surveillance (VS'99)*. IEEE, 14–21.
- OVER, P., AWAD, G., MICHEL, M., FISCUS, J., KRAAL, W., AND SMEATON, A. F. 2011. Trecvid 2011—An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of the TREC Video Retrieval Evaluation (TRECVID'11)*.
- PARK, U. AND JAIN, A. K. 2010. Face matching and retrieval using soft biometrics. *IEEE Trans. Inf. Forensics Security* 5, 3, 406–415.
- PARK, U., JAIN, A. K., KITAHARA, I., KOGURE, K., AND HAGITA, N. 2006. ViSE: Visual search engine using multiple networked cameras. In *Proceedings of the International Conference on Pattern Recognition*. 1204.

- PELLEGRINI, S., ESS, A., SCHINDLER, K., AND VAN GOOL, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*. 261–268.
- PERERA, A. G. A., SRINIVAS, C., HOOGS, A., BROOKSBY, G., AND HU, W. 2006. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition 1*, 666–673.
- PETRUSHIN, V. A., WEI, G., AND GERSHMAN, A. V. 2006. Multiple-camera people localization in an indoor environment. *Knowl. Inf. Syst.* 10, 229–241.
- PETS 2000–2009. Pets: Performance evaluation of tracking and surveillance. <http://www.cvg.rdg.ac.uk/slides/pets.html>.
- PHAM, T. V., WORRING, M., AND SMEULDERS, A. W. 2007. A multi-camera visual surveillance system for tracking of reoccurrences of people. In *Proceedings of the International Conference on Distributed Smart Cameras*. IEEE, 164–169.
- PICCARDI, M. 2004. Background subtraction techniques: A review. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. Vol. 4. 3099–3104.
- PONS-MOLL, G., LEAL-TAIXÉ, L., TRUONG, T., AND ROSENHAHN, B. 2011. Efficient and robust shape matching for model based human motion capture. In *Proceedings of the 33rd International Conference on Pattern Recognition (DAGM'11)*. Springer-Verlag, Berlin, 416–425.
- PORIKLI, F. 2003. Inter-camera color calibration by correlation model function. In *Proceedings of the IEEE International Conference on Image Processing*. Vol. 2. II – 133–6 vol. 3.
- PROJECT, C. 2008. Video and image datasets index. <http://www.hitech-projects.com/euprojects/cantata/>.
- PROSSER, B., GONG, S., AND XIANG, T. 2008. Multi-camera matching under illumination change over time. In *Proceedings of the Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*. Andrea Cavallaro and Hamid Aghajan, Marseille, France.
- PROSSER, B., ZHENG, W., GONG, S., AND XIANG, T. 2010. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference*. 21.1–11.
- RADKE, R. J. 2008. A survey of distributed computer vision algorithms. In *Handbook of Ambient Intelligence and Smart Environments*. H. Aghajan, Ed., Springer.
- REID, D. AND NIXON, M. 2011. Using comparative human descriptions for soft biometrics. In *Proceedings of the 1st International Joint Conference on Biometrics*.
- RIOS-CABRERA, R., TUYTELAARS, T., AND GOOL, L. J. V. 2011. Efficient multi-camera detection, tracking, and identification using a shared set of haar-features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 65–71.
- ROULLOT, E. 2008. A unifying framework for color image calibration. In *Proceedings of the 15th International Conference on Systems, Signals and Image Processing (IWSSIP'08)*, 97–100.
- SALZMANN, M. AND URTASUN, R. 2010. Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'10)*. 647–654.
- SATTA, R., FUMERA, G., AND ROLI, F. 2012a. Fast person re-identification based on dissimilarity representations. *Pattern Recognit. Lett.* 33, 1838–1848.
- SATTA, R., FUMERA, G., AND ROLI, F. 2012b. A general method for appearance-based people search based on textual queries. In *Proceedings of the 1st International ECCV Workshop on Re-Identification (ReID'12)*. Florence, Italy.
- SCHÜGERL, P., SORSCHAG, R., BAILER, W., AND THALLINGER, G. 2007. Object re-detection using SIFT and MPEG-7 color descriptors. *Lecture Notes in Computer Science*, 305–314.
- SCHWARTZ, W. AND DAVIS, L. 2009. Learning discriminative appearance-based models using partial least squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*.
- SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., AND BLAKE, A. 2011. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1297–1304.
- SIVAPALAN, S., CHEN, D., DENMAN, S., SRIDHARAN, S., AND FOOKES, C. 2011. 3d ellipsoid fitting for multi-view gait recognition. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance*. 355–360.
- SMEATON, A. F., OVER, P., AND KRAAIJ, W. 2006. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*. New York, 321–330.
- SONG, B., JENG, T.-Y., STAUDT, E., AND ROY-CHOWDHURY, A. K. 2010. A stochastic graph evolution framework for robust multi-target tracking. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*. Springer-Verlag, Berlin, 605–619.

- TAYLOR, J., SHOTTON, J., SHARP, T., AND FITZGIBBON, A. W. 2012. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 103–110.
- TEIXEIRA, L. F. AND CORTE-REAL, L. 2009. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognit. Lett.* 30, 2, 157–167.
- TUZEL, O., PORIKLI, F., AND MEER, P. 2008. Pedestrian detection via classification on Riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 10, 1713–1727.
- UTSUMI, A. AND TETSUTANI, N. 2004. Human tracking using multiple-camera-based head appearance modeling. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 657–662.
- VAN DE SANDE, K., GEVERS, T., AND SNOEK, C. 2008. Evaluation of color descriptors for object and scene recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1–8.
- VAQUERO, D., FERIS, R., TRAN, D., BROWN, L., HAMPAPUR, A., AND TURK, M. 2009. Attribute-based people search in surveillance environments. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'09)*.
- VELARDO, C. AND DUGELAY, J. 2010. Weight estimation from visual body appearance. In *Proceedings of the 4th IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS'10)*. 1–6.
- VEZZANI, R., BALTIERI, D., AND CUCCHIARA, R. 2009. Pathnodes integration of standalone particle filters for people tracking on distributed surveillance systems. In *Proceedings of the IEEE International Conference on Image Analysis. and Process.* Springer-Verlag, Berlin, Heidelberg, 404–413.
- VEZZANI, R. AND CUCCHIARA, R. 2010. Video surveillance online repository (visor): an integrated framework. *Multimedia Tools Appl.* 50, 2, 359–380.
- VIOLA, P., PLATT, J. C., AND ZHANG, C. 2006. Multiple instance boosting for object detection. In *NIPS 18*. MIT Press, 1419–1426.
- WANG, X.-H. AND LIU, J.-L. 2009. Tracking multiple people under occlusion and across cameras using probabilistic models. *J. Zhejiang University SCIENCE A* 10, 7, 985–996.
- WEBER, M. AND BAUML, M. 2011. Part-based clothing segmentation for person retrieval. In *Proceedings of the IEEE Conference on Advanced Video and Signal-Based Surveillance*. 361–366.
- YANG, B., HUANG, C., AND NEVATIA, R. 2011. Learning affinities and dependencies for multi-target tracking using a crf model. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1233–1240.
- YANG, J., ZHU, X., GROSS, R., KOMINEK, J., PAN, Y., AND WAIBEL, A. 1999. Multimodal people ID for a multimedia meeting browser. In *Proceedings of the International ACM Multimedia Conference*. 159.
- YILMAZ, A., JAVED, O., AND SHAH, M. 2006. Object tracking: A survey. *ACM Comput. Surv.* 38, 4, 13.
- YOON, K., HARWOOD, D., AND DAVIS, L. 2006. Appearance-based person recognition using color/path-length profile. *J. Visual Commun. Image Represent.* 17, 3, 605–622.
- YU, Y., HARWOOD, D., YOON, K., AND DAVIS, L. S. 2007. Human appearance modeling for matching across video sequences. *Mach. Vision Appl.* 18, 3–4, 139–149.
- ZAJDEL, W., ZIVKOVIC, Z., AND KROSE, B. 2005. Keeping track of humans: Have I seen this person before? In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'05)*. 2081–2086.
- ZHANG, L., LI, Y., AND NEVATIA, R. 2008. Global data association for multi-object tracking using network flows. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 1–8.
- ZHENG, W.-S., GONG, S., AND XIANG, T. 2009. Associating groups of people. In *Proceedings of the British Machine Vision Conference*.
- ZHENG, W.-S., GONG, S., AND XIANG, T. 2011. Person re-identification by probabilistic relative distance comparison. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 649–656.
- ZHENG, W.-S., GONG, S., AND XIANG, T. 2012. Re-identification by relative distance comparison. *Pattern Anal. Mach.* 35, 3, 653–6681.
- ZHOU, Q. AND AGGARWAL, J. 2006. Object tracking in an outdoor environment using fusion of features and cameras. *Image Vision Comput.* 24, 11, 1244–1255.
- ZHOU, Y. AND KUMAR, A. 2011. Human identification using palm-vein images. *IEEE Trans. Inf. Forensics Security* 6, 4, 1259–1274.

Received February 2012; revised May 2013; accepted June 2013