A Comparison of Approaches for Person Re-identification

Maria De Marsico[1], Riccardo Distasi[2], Stefano Ricciardi[2], Daniel Riccio[3]
[1]*Dipartimento di Informatica, Sapienza Università di Roma, Rome, Italy*
[2]*Dipartimento di Studi e Ricerche Aziendali, Università di Salerno, Fisciano, Italy*
[3]*Dipartimento di Ingegneria e Tecnologia dell'Informazione, Università di Napoli, Federico II, Napoli, Italy*

Keywords: Re-identification, Biometrics, Multiple Cameras Networks, Persistent Tracking, People Tracking.

Abstract: Advanced surveillance applications often require to re-identify an individual. In the typical context of a camera network, this means to recognize a subject acquired at one location among a feasible set of candidates acquired at different locations and/or times. This task is especially challenging in applications targeted at crowded environments. Face and gait are contactless biometrics which are particularly suited to re-identification, but even "soft" biometrics have been considered to this aim. We present a review of approaches to re-identification, with some characteristic examples in literature. The goal is to provide an estimate of both the state-of-the-art and the potential of such techniques to further improve them and to extend the applicability of re-identification systems.

1 INTRODUCTION

Discontinuous tracking of people across large sites, that is the search of a person of interest in different non-overlapping locations over different camera views, is a crucial task known as people re-identification. A more formal definition of the re-identification paradigm can be summarized as follow: *given a probe set acquired at location X at time T0, re-identification aims to match its items with the subjects in a gallery set, collected at a different location Y at time T1*. Biometrics seems a viable solution to solve this problem when human subjects are to be tracked. It is widely accepted that biometric recognition under controlled data acquisition conditions is a relatively mature technology, which has proved to be effective for the actors of security agencies, public transports, governments, and in independent technology evaluation initiatives (Phillips *et al*., 2010). However, the feasibility of biometric techniques under uncontrolled data acquisition conditions still raises considerable and often well-motivated scepticism. Just for this reason, re-identifying people moving across different sites covered with non-overlapping cameras, could be the ideal scenario for testing technology under uncontrolled data acquisition conditions. As the image data (video) is rather large, the challenge will be to develop fast strategies for human biometric tagging. The application scenarios provide an extremely challenging video analysis task. Video streams will be captured both indoor and outdoor, and the size of people in the images (pixels on target) will range from a few tens to several hundreds pixels. face and gait biometrics are particularly promising for re-identification, since they can operate at a distance and do not require a detailed and/or high resolution image of the subject and/or its biometric traits.

2 PERSON RE-IDENTIFICATION

One of the most critical challenges in facing the person re-identification problem is to recognize the same person viewed by disjoint, possibly non-overlapping cameras, at different time instants and locations. A thorough description of the main issues and open challenges related to remote face identification and face-based re-identification as well, is presented in the work by Chellappa *et al.* (2012) which analyzes the impact of each of the main aspects characterizing unconstrained applicative scenarios, like camera-subject distance, camera resolution, illumination in outdoor environments, pose variation, blur, occlusions and weather artifacts. More recently, a suvery specifically devoted to re-identification has been

De Marsico M., Distasi R., Ricciardi S. and Riccio D..
A Comparison of Approaches for Person Re-identification.
DOI: 10.5220/0004815201890198
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 189-198
ISBN: 978-989-758-018-5

proposed by Vezzani *et al.* (2013). A re-identification technique can be analyzed according to a number of aspects: i) number of tracked subjects; ii) granularity of extracted features (segmented regions, blobs, key-points); iii) possible overlap of camera fields of view; iv) use of contextual information; v) efficiency and effectiveness. In the following we group a number of representative methods according to some relevant criteria, taking into account that a single system may fit more than one category. Table 1, located at the end of this section, resumes all the contributions cited in the subsections 2.1 to 2.6 with regards to authors, methods and datasets used for experiments.

2.1 Single and Multi-target Tracking across Disjoint Cameras

Both the method proposed in (Madden *et al*., 2007), and the one proposed by Colombo *et al.* (2008) are based on a single-person tracking, though exploiting the analysis of very different features. The appearance model proposed by the first one is based on an Incremental Major Colour Spectrum Histogram (IMCHSR). This is obtained by combining an online k-means colour clustering algorithm and the incremental use of frames. The second performs a Gaussian Mixture Model (GMM) segmentation on the video frames, and includes only the foreground data in subsequent steps. The colour constancy between cameras is then processed to improve the quality of segmentation before extracting several person descriptors; these are based on appearance (mean colour, covariance, MPEG-7 dominant colour) and on spatio-temporal properties (Kalman filter-based, topological) of the observations. More recent techniques aim at tracking more objects at the same time, to significantly support video analytics functions, e.g., trajectory recognition, actions and interactions between people and objects. An especially critical aspect is due to the lack of overlap in the camera fields of view, which is usual in most real world applications, where camera locations may even span a large geographic area. As a consequence, acquisition characteristics like illumination, color, or atmospheric conditions, may significantly differ among cameras. This increases the attention devoted to such elements during video processing. In particular, the need for color correction/mapping, for features robust to distortions to characterize persons/objects, often integrated by contextual information, and of optimizations aiming at to increase efficiency. The framework proposed by Truong Congl *et al*. (2010) illustrates an example of the typical strucrture of such architectures. The proposed system consists of two main parts: the first one implements an automatic process for silhouette extraction based on the adaptive GMM in a joint spatial-colorimetric feature space; the second part implements a classification technique, which exploits the discriminative characteristic of sparse representation of signals to perform people re-identification. In (Javed *et al*., 2005), the authors focus on the problems implied by non-overlapping multiple cameras systems, in particular the possibly very different appearance of the same subject due to different camera parameters or lighting conditions. They show that *brightness transfer functions* from a given camera to another camera lie in a low dimensional subspace that can be learned during a training phase and demonstrate that this subspace can be used to compute appearance similarity. To this aim they exploit the Maximum A Posteriori (MAP) estimation framework using both location and appearance cues. In (Javed *et al*., 2008), the authors further extend this approach by observing that in most cases people or vehicles tend to follow the same paths, therefore propose a novel algorithm using this conformity to establish correspondences. By exploiting this property, the algorithm captures the inter-camera relationships in the form of multivariate probability density of space–time variables using kernel density estimation. The space–time and appearance models for object tracking are learned in a training phase. Jeong and Jaynes (2008) estimate the unknown color transfer function between pairs of disjoint cameras by means of a color calibration method. The developed model operates on chromaticity samples to increase the temporal stability of the transfer function between any camera pair. Kuo *et al*. (2010) also address the tracking problem for multiple non-overlapping cameras. In addition, they introduce the association of multi-target tracks by a discriminative appearance affinity model learned on-line. Multiple Instance Learning (MIL) boosting algorithm is adopted to solve the labelling ambiguity during the learning process at runtime. A multi-object correspondence optimization framework is provided to solve the "target handover" problem across cameras.

2.2 The Granularity of Features

An extremely important factor affecting re-identification performance is the type and granularity of the features which are extracted for tracking. In a top-down perspective with respect to granularity, one can classify the different methods according to their basic elements: i) whole objects; ii) segmented objects; iii) key-points; iv) biometric traits. The last kind of elements can be considered as the most fine-grained one, adding a semantic content to the tracked object (face, person). Since it is also strictly bound to the permanence of features used for re-identification, it will be discussed separately.

2.2.1 Methods based on Whole Objects

Among the methods which compute feature vectors starting from the whole detected object we can mention the one proposed by Bak *et al*. (2010a). In this approach, people are first detected and tracked using Histograms of Oriented Gradations (HOG) with a Sobel convolution kernel. The detection uses 15 cells in specific locations around a human silhouette. Then two signatures are obtained from data tracked through a few frames: one is based on Haar-like features combined through a cascade using information entropy as a feature reduction heurystic; the other one is based on Dominant Color Descriptors for the upper and lower part of the body, combined using the AdaBoost scheme. The system is set up with multiple cameras, and the signatures are obtained at one camera and reused at the others. The system performs color normalization across cameras in order to handle color dissimilarities due to differences in lighting or camera calibration. In (Bak *et al*., 2010b), the color-normalized signatures are obtained by computing covariance descriptors on six regions of interest corresponding to fixed body parts. The signatures are matched through a multiresolution grid method. This is a pyramid matching scheme using covariant matrix distance, and modified to include spatial information, where matches found at finer resolutions have more weight than matches found at coarser resolutions. The authors further refine this technique in (Bak *et al*., 2011), where they propose a human appearance signature, called Mean Riemannian Covariance Grid (MRCG). The matrices are seen as tensors on a Riemannan manifold (i.e., a non-Euclidean space without the usual additive structure). The system requires no "learning", but the entrance of a new subject requires updating of all the signatures in the database. The MRCG is also adopted in (Corvée *et al*.) as fusion technique to combine several clues, mainly clothing characteristics, for short term re-identification. People are detected by a simplified Local Binary Pattern operator (SLBP) that extracts a 16-dimensional feature vector submitted to Adaboost training. This method is especially suited for low resolution images where fine features such as iris or even face shots cannot be reliably assumed to be available. The system can also withstand some changes in clothing (such as unzipping a jacket). Ayedi *et al*. (2012) also focus their attention on the relevance of adequate descriptors for representation and ultimately for re-identification purposes in the context of camera networks. Uncontrolled settings relative to the field of view (FOV) available to each camera node require relative invariance.

2.2.2 Methods based on Segmented Objects

A first step towards a finer granularity of features for tracking of objects is the introduction of texture features, as proposed in (Berdugo *et al*., 2010). The integration of textural features, alongside of the conventional color features, and a probabilistic model of a human, provides a measurable improvement in the correct matching of similar figures and therefore an increment in the re-identification success rate. The metric exploited to evaluate the correlation between two appearance models is the Kullback-Leibler distance, matching key frames from the trajectory path of the object, rather than matching a single image. Farenzena *et al*. (2010) exploit textural information by detecting the presence of recurrent local motifs with high entropy. They combine textures with the overall chromatic content and the spatial arrangement of colours into stable regions. The main idea is to extract features that model three complementary aspects of the human appearance. Some techniques, like the one proposed in (Chien *et al*., 2006), partition the human silhouette in regions from which local features are extracted. The authors propose a multi-stage algorithm based on a human-specific descriptor referred as Human Color Structure Descriptor or HSCD, representing the colors of body, legs, and shoes of a human object at specific positions in a compact 112 bits vector. The main aim of this approach is the integration of fundamental tasks of re-identification, namely segmentation, tracking and description generation. Indeed, it is able to achieve better performance than Scalable Color Descriptor and Color Structure Descriptor while requiring a reduced computation. Wang *et al*. (2007) also exploit several statistics over image subregions modelling, related to both shape and appearance context. The descriptor of a given object is its occurrence matrix resulting from a novel fast real-time algorithm to find occurrence and co-occurrence, which is based on integral computations to increase the algorithm efficiency. Correspondence between multiple cameras for visual surveillance topic is addressed by Hu *et al.* (2006), by introducing a technique based on principal axes of people, where the people similarity match across multiple cameras exploits the relationship between "ground-points" detected in each view and the intersections of the principal axes detected in different views and transformed to the same view. In the context of approaches using local features it is also to consider methods based on the *bag-of-features.* In (Gray and Tao, 2008) the bag-of-features are used to provide a technique to perform viewpoint invariant pedestrian recognition adopting the Ensemble of Localized Features (ELF) as object representation. The ELF model contains 200 color and texture-based features processed by Schmid and Gabor filters; an AdaBoost based function computes similarity. The pedestrian recognition task requires that the ELF model is supervised by a human operator; on the other hand, it is effective and totally automatic at discriminating between pedestrians regardless of the viewpoint change. Doretto *et al.* (2011) review and compare several aspects of appearance-based techniques for person re-identification and illustrate two local descriptors in more depth: the histogram of oriented gradients in log-color space (HOG log-RGB) and the HVS edgel technique, along with a detailed description of the salient edgel extraction algorithm. Organization of this lower-level data relies on bounding box and bag-of-features models. As for signature computation, the paper describes a quick method to compute co-occurrence matrices exploiting integral image representation. Covered parts-based models include interest point matching and model fitting.

2.2.3 Some Comparison of Local Approaches

Given the high number of local features available, it is crucial to analyse which ones can provide a good support for person-reidentification. Bäuml and Stiefelhagen (2011) compare several local features for re-identification in image sequences. Assuming that a tracker module provides a rough bounding box around the picture of a person, features are computed for all interest points lying within the bounding box. The model applied is that of a bag of features, meaning that spatial and temporal information about the features is discarded. The interest point detectors evaluated are Harris, Harris-Laplace, Hessian-Laplace, Harris-affine, Hessian-affine, and Fast-Hessian. The local descriptors used are SIFT, Shape Context (SC) with Canny edge detector, Gradient Location and Orientation Histogram (GLOH), and SURF. Since video footage is actually acquired rather than still images, feature distances are fused with the sum rule across frames. The experimental results indicate that no specific point detector has a definite advantage over the others. Among the descriptors, GLOH and SIFT performed better than the other two. A further interesting evaluation regards the comparative analysis between local features and more articulated and complex characterization tools, such as the edgel deeply reviewed by Doretto *et al.* (2011). This is the aim of the sudy by Gheissari *et al.* (2006), where the authors compare two methods for people re-ideinfication. The first one is based on points of interest obtained via the Hessian affine invariant operator, which has the advantage of generating more data points where the information content is high. The second method is a dynamic model for the time-variant appearance of a subject based on edgel extraction. The model is obtained from a decomposable triangulated graph, fitted to the images by a dynamic programming algorithm. The approach based on points of interest uses local color histograms for comparison and is computationally lighter, but results do not persist over time due to the changing appearance of subjects. The dynamic model maps body parts from subject to subject so that a correspondence can be established through matching scores. The authors also propose a spatio-temporal segmentation algorithm that is robust to variations in appearance of clothes due to folds and wrinkles. The two methods are compared along with a baseline reference method based on the foreground histogram of a crude bounding box. The dynamic model is found to perform significantly better than the other two, whose performance is similar at the experimental resolution. According to the results of these comparative studies, it appears that fusion of information from different levels is more advantageous that selection to improve the efficacy of person re-identification techniques.. As a matter of fact, Bazzani *et al.* (2010) propose an appearance-based method based on a signature called HPE (Histogram Plus Epitome), which incorporates both local and global descriptors. The emphasis is on the overall chromatic content, but there are additions that account for local patches of color. The input is a sequence of single-camera still images. The background is eliminated by the STEL generative model, then the HSV histograms of the foreground undergo an unsupervised Gaussian clustering. The HSV histogram is the first component of the signature and accounts for global color distribution.

The signature has two more components (epitomes) that describe the color patches present in the subject at the global and local level. The experimental results show the importance of accurate calibration in the training set: increasing the number of images used to compute the descriptors has significantly diminishing returns. From one side it is possible to select and combine more types of features. From the other side, it is possible to work on the distance measures adopted to compare such features. In particular, it is possible to have such distance dynamically adapt in time to the changing context where it is used. This is the core idea of the work by Zheng *et al*. (2011), whose aim is that of building a learning system that can iteratively find an optimal distance function that maximizes the probability of having a true matching pair lie at a smaller distance than a mismatching pair, with the maximization performed over all possible matching and mismatching pairs. Matching can use raw image data or extracted features such as rough histograms. The distance function used as a starting point for learning include color histograms from horizontal stripes of a person's image as well as texture information extracted by Schmid and Gabor filters.

2.2.4 Methods based on Key-points

Going down along the granularity rank of the features exploited for re-identification, we find the techniques based on key-points produced by transforms (e.g., Scale Invariant Feature Transform – SIFT, Speeded Up Robust Features – SURF). Jungling and Arens (2010) adopt SIFT features for person re-identification in infrared image sequences. After a clustering stage where feature prototypes are built, an Implicit Shape Model (ISM) records the spatial occurrence of features in terms of object center offsets. The person re-identification module adopts a novel model that uses the general appearance codebook applied for person detection as an indexing structure for re-identification, and thus is able to efficiently acquire and match models. To this aim SIFT features collected during training are integrated into the person instance model; the latter is indexed by the codebook entries which activated the feature during detection. The module for matching models uses a two level strategy: the first stage allows for fast discovery of promising models based on matching of person signatures, while the second one performs a detailed analysis of models based on feature descriptors. Oliveira and Luiz (2009) propose a different approach based on local features. Interest points collected in a query image are matched with those collected in each video sequence used for each previously seen person. The system exploits multiple networked cameras featuring local processing, via an onboard processor running Linux OS. This design may represent a limitation and/or a complication, but it shows a good accuracy in the re-identification even in presence of cloth changes or occlusions. Hamdoun *et al*. (2008) propose a method based on interest points given by SURF (Hessian and Haar wavelets). The interest points are extracted from video sequences as opposed to single isolated frames. Query data are also gathered from video sequences, yet shorter than those in the learning phase. The metric in descriptor space is the sum of absolute distances, and the matching relies on best-bin-first search in a KD-tree with all models. Query/model matching is performed by multiple interest points voting for a particular match. The variant of SURF used aims at maximum time efficiency and only uses integer arithmetic.

2.3 Permanence of Biometric Features

An important aspect that many techniques described so far do not consider is time. They are effective to re-identify the same person in videos produced by disjoint cameras, but in a reduced time window. When the time elapse increases (i.e. more days) it is not realistic to assume that a person wears the same dresses, so that in this case many methods risk to be completely ineffective. On the contrary, methods based on biometric traits can also be applied in contexts of this kind. For this reason such methods are quickly spreading. In the context of person re-identification, face appears to offer the best compromise between concrete feasibility of the systems and accuracy of the recognition: This is due to the fact that the face surface is surely larger than other physical biometric traits, e.g., the iris, and that it requires a contact-less acquisition, differently from, e.g., fingerprints. At the same time it guarantees a reasonable accuracy even in under-controlled conditions. It is to consider in any case that face recognition in unconstrained settings, including unstable data capture conditions, is very challenging. As a matter of fact, many algorithms perform well, when constrained face images are acquired, but their performance degrades significantly when the test images contain variations that are not present in the training images. Chellapa *et al.* (2012) highlight some of the key issues in remote face recognition, and introduce a remote face database which has been acquired in an unconstrained outdoor maritime environment. The problem of face re-identification by disjoint cameras is deeply investigated in Bäuml *et al.* (2010). Face detection is based on the modified consensus transform and a subset of the frames is scanned for different possible face orientations, accounting for possible inter-subject occlusion. User feedback is used to help the subject-specific classifier by pointing out the best and the worst candidate matches. Fisher *et al.* (2011) further extended this work but readapt it to a different application, that is a video retrieval framework, in which the system finds occurrences of a query person in a set of TV episodes. Because of either the limited resolution of acquisition devices or the large distance of the target from them, soft biometrics traits, like hair, skin and cloth patches can significantly support the face re-identification process, as proposed by Dantcheva and Dugelay (2011). In this paper the main aim of the authors is to address pose variations related to the camera network of video-surveillance systems. This is one of the typical challenges of face recognition, and is complicated in this context by an unattended acquisition procedure. The idea is to mimic the way in which humans improve the accuracy of frontal-to-side recognition by exploiting simple and evident traits subdivided into trait-instances, e.g. blond, brown, red and black for the trait hair-color. The proposed algorithm analyzes color and texture of the selected patches, then a combined classifier is built to boost and combine all considered traits. Results do not qualify this soft-biometrics based approach as a candidate for robust re-identification, but rather as a pruning system for high security solutions for access control or as part of a multi-biometrics systems. A further (behavioural) biometrics which can be used for person re-identification is gait analysis, since it can work at a distance even at low resolutions and without co-operation. Roy *et al.* (2012), propose a hierarchical framework for re-identifying a non-cooperative subject by combining gait with the phase of motion in a spatiotemporal model. They use three features: two for subject's motion dynamics, and the third is derived from the spatio-temporal model of the camera network. By combining them, the method can track subjects even if they change speed, or stop for some time in the blind gap.

2.4 Overlapping Fields of View

Though often unaffordable for obvious practical economy reasons, multiple overlapping cameras can provide both more robust detection and 3D location estimation of objects, by covering object features from all directions. When the system handles a model of the overlap, the overlapping vision fields can also support a smooth switching among cameras during tracking. In the work by Cai and Aggarwal (1999) tracking continues from a single camera view until the system predicts that the active camera is going to loose a good view of the subject of interest. At that time tracking switches to the camera that is estimated to provide a better view and that requires the least switching. Three basic modules are involved: Single View Tracking (SVT), Multiple View Transition Tracking (MVTT), and Automatic Camera Switching (ACS). During SVT, a Bayesian classifier locate the most likely match of the subject in the next frame. MVTT involves both spatial and temporal motion estimation. Finally, ACS relies on a prediction algorithm. Mittal and Davis (2003) propose a multi-camera person tracking system based on a region-based stereo algorithm that finds 3D points inside an object from information about regions belonging to the object obtained from two different views. Association across frames exploits the color models of the horizontal sections of the person , which can also be used for reidentification over longer time intervals. Gandhi and Trivedi (2007) develop the concept of Panoramic Appearance Map (PAM) for person re-identification in a multi-camera setup. Each person is tracked in multiple cameras and the position on the floor plan is determined using triangulation among different views. Using the geometry of the cameras and the person location, the system creates a panoramic map centred at the person's location. In the map the horizontal axis represents the azimuth angle and the vertical axis representing the height. If the person is detected in more than one camera, the floor position can be obtained by finding the intersection of the corresponding rays. In order to ensure an accurate projection, only the frames where three or more cameras detect the object are used.

2.5 Example Use of Context

A critical challenges in person re-identification is to recognize the same person viewed by disjoint cameras at different times and locations. Some techniques, e.g., those in (Sankaranarayanan *et al.*, 2008) and (Mazzon *et al.*, 2012), integrate information from feature tracking with those related to context or environment. Mazzon *et al.* (2012) link the appearance of people, the spatial location of cameras, and the potential paths a person can choose to follow. The technique adopts a Landmark-Based Model (LBM) using people movements in non-observed regions, a site map, and regions of interest where people are likely to transit. Sankaranarayanan *et al*. (2008) particularly highlight the efficient use of the geometric constraints induced by the imaging devices, to derive distributed algorithms for target detection, tracking, and recognition. Their conclusions underline the importance of an integrated approach for solving the interdisciplinary problems involved in automated monitoring.

2.6 A Note about Efficiency

An open problem for re-identification is computational cost, especially with camera networks for real-time applications. Satta *et al*. (2012) use the Multiple Component Dissimilarity (MCD) framework, where an appearance-based method leverages dissimilarity-based distances. As the dissimilarity-based descriptors are vectors of real numbers, they are much more compact and the time for matching them is greatly reduced. Moreover, the dissimilarity-based representation used by MCD allows very fast re-identification implementations.

3 REFERENCE DATASETS

The datasets used for testing and fine-tuning are critical for the design, development and performance

Table 1: Resumes of contributions to the field of person re-identification referenced throughout the text.

#	Authors	Dataset	Approach	#	Authors	Dataset	Approach
1	Ayedi et alii	VIPeR	region covariance descriptor based on multi-scale features	19	Gray et alii	VIPeR	Ensemble of Localized Features (ELF)
2	Bak et alii	CAVIAR, TREC	Haar-like features and DCD, HOG based detection and tracking	20	Hu et alii	NLPR	body principal axis
3	Bak et alii	i-LIDS	HOG tracking, covariance descriptors with color normalization, spatial pyramid matching	21	Hamdoun et alii	CAVIAR	regions of interest: Haar/SURF variant.
4	Bak et alii	i-LIDS, ETHZ	covariance descriptors with custom fusion based on Riemann geometry.	22	Javed et alii	Proprietary	multivariate probability density of space-time variables
5	Bauml and Stiefelhagen	CAVIAR, Proprietary subset	points of interest, local descriptors	23	Javed et alii	Proprietary	training-based, MAP estimation of location and appearance cues
6	Bauml et alii	Proprietary	multiple detectors based on Modified Census transform	24	Jeong and Jaynes	Proprietary (Terrascope)	estimation of unknown color transfer function between pairs of cameras
7	Bazzani et alii	i-LIDS, ETHZ	HSV histograms of global/local color content	25	Jungling and Arens	CASIA	SIFT features, Implicit Shape Model (ISM),
8	Berdugo et alii	GBSEO	appearance modeling, human probabilistic model	26	Kuo et alii	Proprietary	learned discriminative appearance affinity model
9	Cai and Aggarwal	Proprietary	Bayesian classification schemes, multivariate normal distribution	27	Madden et alii	Proprietary	Major Colour Spectrum Histogram Representation (MCSHR)
10	Chellappa et alii	Proprietary	PCA+LDA+SVM, Sparse Representation-based Classification (SRC)	28	Mazzon et alii	i-LIDS	Landmark Based Model (LBM)
11	Chien et alii	Proprietary	Human Color Structure Descriptor (HCSD)	29	Mittal and Davis	Proprietary	Region-based stereo algorithm, bayesian classification
12	Colombo et alii	Proprietary	Gaussian Mixture Model (GMM)	30	Oliveira and Luiz	CAVIAR	Hessian matrix, Haar wavelet
13	Corvée et alii	INRIA, proprietary	LBP, ADAboost, Mean Riemannian Covariance Grid (MRCG)	31	Roy et alii	Proprietary	gait, phase of motion, spatiotemporal model
14	Dantcheva and Dugelay	FERET	AdaBoost applied to hair, skin and clothes patches	32	Sankaranarayanan et alii	VIPeR	Multiple Component Matching (MCM) based on Multiple Instance Learning
15	Doretto et alii	Proprietary	comparison of several approaches	33	Satta et alii	VIPeR, i-LIDS	Multiple Component Dissimilarity (MCD)
16	Farenzena et alii	VIPeR, i-LIDS, ETHZ	HSV histogram, Maximally Stable Colour Regions (MSCR), recurrent highly structured patches	34	Truong Cong et alii	Proprietary	color-position histogram, spectral analysis, SVM
17	Gandhi and Trivedi	Proprietary	multi-camera based Panoramic Appearance Map (PAM)	35	Wang et alii	Proprietary	shape and appearance context modeling
18	Gheissari et alii	Proprietary	Hessian-affine regions of interest, dynamic graph-based model	36	Zheng et alii	VIPeR, i-LIDS	dynamic definition of a custom distance function.

assessment of a re-identification system. Their content may greatly affect the results from each system, e.g., due to resolution, noise, illumination conditions, camera distance from the subjects, average number of people in each frame, etc. Some publicly available dataset are briefly described, while details can be found at the corresponding sites.

VIPeR (Viewpoint Invariant Pedestrian Recognition): 632 images (128x48 pixels) taken from arbitrary viewpoints under varying conditions. http://vision.soe.ucsc.edu/?q=node/178

CAVIAR (Context Aware Vision using Image-based Active Recognition): Video clips (384 x 288 pixels, 25 fps) of people walking alone, meeting with others, window shopping, etc. http://homepages.inf.ed.ac.uk/rbf/CAVIAR.

TREC (Video Retrieval Evaluation Data): ~100-hour corpus of video from 5 cameras**;** contains surveillance data collected by the UK Home Office at the London Gatwick International Airport**.** http://www.itl.nist.gov/iad/mig//tests/trecvid/2008/doc/EventDet08-EvalPlan-v06.htm.

INRIA (Person Dataset): images with different characteristics aimed to detection of upright people**.** http://pascal.inrialpes.fr/data/human/.

i-LIDS (Imagery Library for Intelligent Detection Systems): library of CCTV video footage from 5 cameras; dataset versions containing still images are also available; event detection scenarios include sterile zone, parked vehicle, abandoned baggage, doorway surveillance; a tracking scenario with multiple camera is also included**.** https://www.gov.uk/imagery-library-for-intelligent-detection-systems.

CASIA (Gait Database): image sequences (Dataset A) and videos. Dataset A, Dataset B (multiview dataset – 11 views), Dataset C (infrared dataset), Dataset D (real surveillance scenes). http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp.

GBSEO: video streams (one hour long, split in 5 minutes clips) by two (not synchronized) cameras at the same time from different viewpoints with some overlap. Different people that appear and reappear few times in both cameras - videos are manually annotated with bounding boxed and person names**.** http://sipl.technion.ac.il/GBSEO.shtml.

ETHZ: samples from street video sequences. http://homepages.dcc.ufmg.br/~william/datasets.html

PETS 2009 (Benchmark data for Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance): video footage from 8 cameras (7 fps) containing different crowd activities. http://www.cvg.rdg.ac.uk/PETS2009/a.html.

Videoweb Activities Dataset – **Items:** 2.5 hours of video from 8 cameras (30 fps) in courtyard. http://www.ee.ucr.edu/~amitrc/vwdata.php.

3D PES (3D People Surveillance Dataset): video sequences from 8 video surveillance cameras**.** http://www.openvisor.org/3dpes.asp

4 CONCLUSIONS

The efficiency of a person re-identification system is obviously bound to the application. Real-time processing imposes severe constraints. Machine learning algorithms may require a kind of training that may not be feasible in highly dynamic settings. The use of one or more cameras, and of overlapping or non-overlapping fields of view, has its implications too. Single camera applications are nowadays limited to simple settings. On the other hand, while overlapping views provide more robust information and allow more "intelligent" processing, it is also to consider that possible triangulation is computationally demanding. Therefore, the choice of the system to use is bound to the real operational needs. A plurality of issues must be addressed when facing re-identification problems, including clustering and selection of relevant regions, recognition-by-parts, anomaly and change detection, sampling and tracking, fast indexing and search, sensitivity analysis, and their ultimate integration. This implies plenty of sparks for present and future research. Many problems are borrowed from complementary fields, such as biometric recognition, so that also solutions can be adapted and improved. The challenge addressed is an evidence-based management to progressively collect and add useful information to data, in order to generate knowledge and appropriate triggering of pre-determined actions.

REFERENCES

Ayedi W., Snoussi H., Abid M., 2012. A fast multi-scale covariance descriptor for object re-identification. In *Pattern Recognition Letters*, 33(14), pp. 1902-1907.

Bak S., Corvee E., Bremond F., and Thonnat M., 2010a. Person Re-identification Using Haarbased and DCD-based Signature. In *2nd AMMCSS*, pp. 1-8.

Bak S., Corvee E., Bremond F., and Thonnat M., 2010b. Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In 7th *Intl. Conf. on Advanced Video and Signal-Based Surveillance,* pp. 435-440.

Bak S., Corvee E., Bremond F., and Thonnat M., 2011.

Multiple-shot human re-identification by mean Riemannian covariance grid. *In 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pp. 179-184.

Bäuml M., Bernardin K., Fischer M.,. Ekenel H. K, 2010. Multi-pose face recognition for person retrieval in camera networks. In *IEEE 7th Int. Conf. on Advanced Video and Signal Based Surveillance*, pp. 441-447.

Bäuml M., Stiefelhagen R., 2011. Evaluation of local features for person re-identification in image sequences. In 8th *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pp. 291-296.

Bazzani L., Cristani M., Perina A., Farenzena M., and Murino V., 2010. Multiple-Shot Person Re-identification by HPE Signature. *In Intl. Conf. on Pattern Recognition*, pp. 1413–1416.

Berdugo G., Soceanu O., Moshe Y., Rudoy D., and Dvir I., 2010. Object re-identification in real world scenarios across multiple non-overlapping cameras. In *European Signal Processing Conf.*, pp. 1806-1810.

Cai Q. and Aggarwal J.K., 1999. Tracking Human Motion Using a Distributed-Camera System. In *IEEE Trans. On PAMI.*, 21(12), pp. 1241-1247.

Chellappa R., Ni J., Patel V. M., 2012. Remote identification of faces: Problems, prospects, and progress. In *Pattern Recognition Letters*, 33(14), pp. 1849-1859.

Chien S.Y., Chan W.K., Cherng D.C., Chang J.Y., 2006. Human object tracking algorithm with human color structure descriptor for video surveillance systems. In *IEEE Int. Conference on Multimedia and Expo*, pp. 2097-2100.

Colombo A., Orwell J., and Velastin S., 2008. Colour constan-cy techniques for re-recognition of pedestrians from mul-tiple surveillance cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008, Marseille, France.

Corvée E., Bak S., and Brémond F., 2012. People Detection and Re-identification for Multi Surveillance Camera. In *VISAPP 1*, SciTePress, pp. 82-88.

Dantcheva A.,. Dugelay J. L, 2011. Frontal-to-side face re-identification based on hair, skin and clothes patches. In *IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance*, pp. 309-313.

Doretto G., Sebastian T., Tu P., Rittscher J., 2011. Appearance-based person reidentification in camera networks: Problem overview and current aspects. In *Journal of Ambient Intelligence and Human Computing*, 2(2), pp. 1–25.

Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani M., 2010. Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2360-2367.

Gandhi T.and Trivedi M. M., 2007. Person tracking and reidentification: Introducing panoramic appearance map (PAM) for feature representation. In *Machine Vision Applications*, 18(3), pp.207–220,

Gheissari N., Sebastian T. B., Tu P. H., Rittscher J., and Hartley R., 2006. Person Reidentification Using SpatioTemporal Appearance. In *IEEE Conf. on Comp. Vision and Pattern Rec.,* vol. 2, pp. 1528–1535.

Gray D., Tao H., 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Conf. on Computer Vision*, pp. 262-275.

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 *ACM/IEEE Intl. Conf. on Distributed Smart Cameras),* pp.1–6.

Hu W., Hu M., Zhou X., Lou J., 2006. Principal Axis-Based Correspondence between Multiple Cameras for People Tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), pp.663-671.

Javed O., Shafique K., Shah M., 2005. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conf. on Computer Vision and Pattern Recognition,* vol. 2, pp. 26-33.

Javed O., Shafique K., Rasheed Z., Shah M., 2008. Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding* 109(2), pp.146–162.

Jeong K., Jaynes C., 2008. Object matching in disjoint cameras using a colour transfer approach. *Journal of Machine Vision and Applications*, 19(5), pp. 88–96.

Jungling K., Arens M., 2010. Local Feature Based Person Re-identification in Infrared Image Sequences. In 7° *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 448-455.

Kuo C. H., Huang C., Nevatia R., 2010. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *Conference on Computer Vision*, pp. 383-396.

Madden C., Cheng E. D., Piccardi M., 2007. Tracking people across disjoint camera views by an illumination tolerant appearance representation. *Journal of Machine Vision and Applications* 18(3), pp. 233–247.

Mazzon R., Tahir S. F., Cavallaro A., 2012. Person re-identification in crowd. In *Pattern Recognition Letters*, 33(14), pp. 1828-1837.

Mittal A., Davis, L. 2003. M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene*. Int. J. Comput. Vis*. 51(3), pp.189–203.

Oliveira I., Luiz J., 2009. People re-identification in a camera network. In *IEEE Int. Conf. on Dependable, Autonomic and Secure Computing*, pp. 461-466.

Phillips P. J., Scruggs T., O'Toole A., Flynn P. J., Bowyer K. W., Schott C. and Sharpe M., 2010. FRVT 2006 and ICE 2006 Large-Scale Experimental Results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5), pp. 831–846.

Roy A., Sural S., Mukherjee J., 2012. A hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification. In *Pattern Recognition Letters*, 33(14), pp. 1891-1901..

Sankaranarayanan A. C., Veeraraghavan A., and Chellappa R., 2008. Object detection, tracking and recognition for multiple smart cameras. In

Proceedings of the IEEE, vol.96, no.10, pp.1606-1624.

Satta R., Fumera G., Roli F., 2012. Fast person re-identification based on dissimilarity representations. In *Pattern Recognition Letters*, 33(14), pp. 1838-1848.

Truong Congl D.-N., Khoudour L., Achard C., Meurie C., Lezoray O., 2010. People re-identification by spectral classification of silhouettes. *Signal Processing*, vol. 90, no. 8, pp 2362-2374.

Vezzani R., Baltieri D., Cucchiara R.,2013. People Re-identification in Surveillance and Forensics: a Survey. Accepted for publication in ACM Computing Surveys.

Wang X., Doretto G., Sebastian T., Rittscher J., Tu P., 2007. Shape and appearance context modeling. In *IEEE Int. Conf. on Computer Vision*, pp. 1-8.

Zheng W., Gong S., and Xiang T., 2011. Person re-identication by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 649-656.