Person re-identification in crowd

Riccardo Mazzon*, Syed Fahad Tahir, Andrea Cavallaro

Queen Mary University of London,
School of Electronic Engineering and Computer Science,
Mile End Road, London E1 4NS, UK

Abstract

Person re-identification aims to recognize the same person viewed by disjoint cameras at different time instants and locations. In this paper, after an extensive review of state-of-the-art approaches, we propose a re-identification method that takes into account the appearance of people, the spatial location of cameras and potential paths a person can choose to follow. This choice is modeled with a set of areas of interest (landmarks) that constrain the propagation of people trajectories in non-observed regions between the field-of-view of cameras. We represent people with a selective patch around their upper body to work in crowded scenes when occlusions are frequent. We demonstrate the proposed method in a challenging scenario from London Gatwick airport and compare it to well-known person re-identification methods, highlighting their strengths and limitations. Finally, we show by Cumulative Matching Characteristic curve that the best performance results by modeling people movements in non-observed regions combined with appearance methods, achieving an average improvement of 6% when only appearance is used and 15% when only motion is used for the association of people across cameras.

Keywords: Person re-identification, non-overlapping cameras, people in crowd, trajectory propagation, appearance features, London Gatwick airport dataset

1. Introduction

The surveillance of wide areas such as airports and train stations requires the deployment of networks of cameras whose field-of-views (FOV) may be disjoint. Disjoint cameras make person re-identification a challenging problem, because of changes in pose, scale and illumination that modify the perceived appearance of a person across cameras (Fig. 1). Moreover, in a crowd, the full body is often not visible due to occlusions. Finally, people exiting the FOV of a camera may enter in different regions of the FOV of the next camera so the time needed to travel across cameras and the area of reappearance are variable and difficult to model.

We can identify four main phases in person re-identification, namely multi-person detection, feature extraction, cross-camera calibration, and person association (Fig. 2). The first phase, multi-person *detection*, extracts image regions corresponding to people (Enzweiler and Gavrila, 2009), based on a trained classifier, a motion detector or a combination of both. The second phase extracts *features* from the detected people. Appearance features include color, texture and shape, which can be used separately or combined (Gray and Tao, 2008). These features can be extracted from a single snapshot of the target (Zheng et al., 2011) or, when intra-camera tracking information is available, after grouping features over time (Berdugo et al., 2010). The third phase, *cross-camera calibration*, establishes the color and spatio-temporal relationship across cameras and allows to account for the variability of observations of the same person across different FOVs. Spatio-temporal calibration methods encapsulate information about the camera deployment, the spatial relationship between cameras, the entry/exit points in the scene, and the traveling time across cameras (Javed et al., 2008). Finally, the *association* of candidates across cameras matches different instances of the same person using the information extracted in the previous phases.

*Corresponding author
Email address: riccardo.mazzon@eecs.qmul.ac.uk (Riccardo Mazzon)
URL: http://www.eecs.qmul.ac.uk/~andrea/ (Andrea Cavallaro)

To appear in Pattern Recognition Letters, 2012

Figure 1: Change of people appearance across cameras (iLIDS, 2008). Column 1: camera 1, full frame; Column 2: corresponding crop of a person of interest; Column 3: camera 2, full frame; Column 4: corresponding crop of a person of interest. People appear under different illumination conditions, as shown in (b) and (d), and under different poses and levels of occlusion, as shown in (f) and (h).

Existing person re-identification methods are validated on snapshot-based or video-based datasets. VIPeR (Farenzena et al., 2010; Prosser et al., 2010) and i-LIDS-static (Farenzena et al., 2010; Bak et al., 2010; Prosser et al., 2010; Zheng et al., 2011) are the most common snapshot-based datasets used to validate appearance-based methods mostly containing people with full body visibility. VIPeR consists of 632 images taken from two outdoor views (Gray and Tao, 2008), while i-LIDS-static contains from 44 (Bak et al., 2010) to 479 (Zheng et al., 2011) image pairs of people taken from four cameras at London Gatwick airport. A video-based dataset is the *Terrascope* dataset (Jeong and Jaynes, 2008) that consists of nine indoor cameras where eight people walk and act in an office environment. Javed *et al.* (Javed et al., 2008) presented a video-based dataset with three sequences composed of up to three cameras from indoor and outdoor scenarios with large illumination changes and up to four fully visible people. Finally, a more challenging dataset in terms of occlusions is that with three outdoor cameras where up to ten people walk alone or in small groups (Kuo et al., 2010).

In this paper, we present a unifying overall structure and an in-depth survey of the state-of-the-art for person re-identification methods that allow us to identify the major common features and drawbacks of existing approaches. Unlike previous method-based surveys (Doretto et al., 2011), our survey has a phase-based organization. Based on the outcome of this survey, we propose a method that (i) integrates simple knowledge of the site under surveillance, (ii) models people movements in non-observed regions using landmark points (regions of interest) in the scene, and (iii) can cope with crowded scenes. The association method uses distances based on appearance, location, and their combination. Appearance features are extracted from a selected area of the upper body (the most visible in case of crowd) and candidate locations are generated using landmark points and people motion. We compare the most representative state-of-the-art person re-identification methods and our proposed method on the London Gatwick airport dataset (iLIDS, 2008) by Cumulative Matching Characteristic (CMC) curves (Gray and Tao, 2008; Prosser et al., 2010; Zheng et al., 2011).

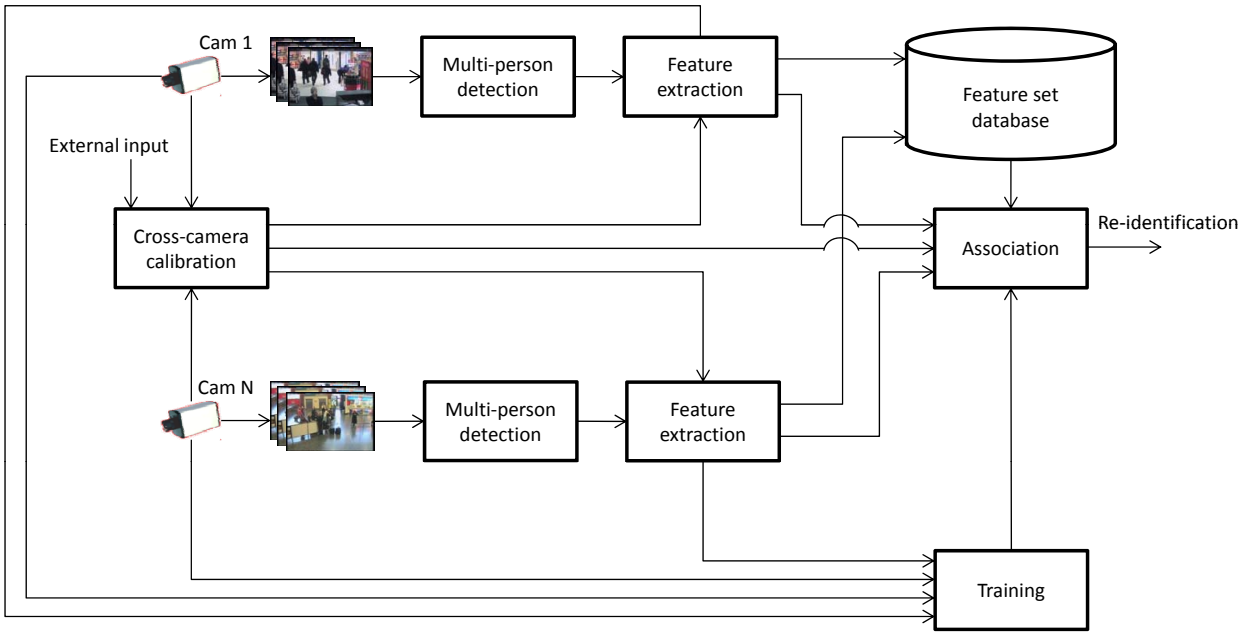


Figure 2: Unifying block diagram for person re-identification approaches.

The paper is organized as follows. Section 2 discusses a comprehensive survey on person re-identification methods, based on their four main phases. Section 3 presents our framework for re-identification that uses a landmark-based spatio-temporal modeling and an upper body representation. In Sec. 4, we validate the proposed approach and compare it with the most common state-of-the-art methods. Finally, Sec. 5 discusses the results and draws conclusions.

2. Person re-identification: a survey

In this section, we discuss person re-identification methods presented in the literature based on the classification in phases proposed in the previous section. The methods are summarized in Table 1. In the following, we consider the first phase (multi-person detection) to have been already solved.

2.1. Features

Color, texture and shape are the appearance features commonly used in the state-of-the-art methods for person re-identification, where features are usually combined in order to obtain a more representative descriptor of the target. Furthermore, temporal consistency of features can be exploited to merge the available information of a person over time.

Color is the most commonly used appearance feature encoded in the form of either histograms (Gheissari et al., 2006; Gray and Tao, 2008; Javed et al., 2008; Oliveira and Luiz, 2009; Cheng et al., 2009; Farenzena et al., 2010; Prosser et al., 2010; Kuo et al., 2010; Zheng et al., 2011) or cumulative histograms (Berdugo et al., 2010), which are simple to compute and scale invariant. Different color channels and their combination can be used: the Hue channel from the HSV color space (Oliveira and Luiz, 2009); the Hue and Saturation channels jointly (Gheissari et al., 2006); or the three channels of the HSV color space (Farenzena et al., 2010). Also, the histogram of the RGB color space is widely used (Prosser et al., 2008; Javed et al., 2008; Cheng et al., 2009; Berdugo et al., 2010). A concatenation of histograms from RGB, YCbCr, and HS (from HSV) color channels (Fig. 3) is adopted in (Gray and Tao, 2008; Prosser et al., 2010; Zheng et al., 2011). An analysis by boosting classifier (Gray and Tao, 2008) shows how, for the re-identification task, the Hue channel is the most discriminative followed by Saturation, Blue, Red, and Green channels. However this analysis is limited to scenes where people are fully visible. Alternatively, the two chrominance

Table 1: State-of-the-art methods for person re-identification. Legend: Spatio-temp = Spatio-temporal, Distance = Distance based, Learning = Learning based, Optim = Optimization based.

Ref.	Appearance features			Temporal grouping	Calibration		Association		
	Color	Texture	Shape		Color	Spatio-temp	Distance	Learning	Optim
(Porikli, 2003)	✓				✓				✓
(Gheissari et al., 2006)	✓	✓		✓			✓		
(Madden et al., 2007)	✓			✓			✓		
(Wang et al., 2007)	✓	✓	✓					✓	
(Hamdoun et al., 2008)		✓		✓			✓		
(Prosser et al., 2008)	✓				✓	✓	✓		
(Jeong and Jaynes, 2008)	✓				✓		✓		
(Gray and Tao, 2008)	✓	✓						✓	
(Javed et al., 2008)	✓			✓	✓	✓			✓
(Oliveira and Luiz, 2009)	✓	✓					✓		
(Teixeira and Corte-Real, 2009)		✓		✓				✓	
(Cheng et al., 2009)	✓	✓	✓			✓			✓
(Farenzena et al., 2010)	✓	✓		✓			✓		
(Bak et al., 2010)	✓	✓		✓			✓		
(Berdugo et al., 2010)	✓	✓	✓	✓			✓		
(Prosser et al., 2010)	✓	✓						✓	
(Bauml et al., 2010)		✓						✓	
(Kuo et al., 2010)	✓	✓	✓	✓		✓			✓
(Zheng et al., 2011)	✓	✓						✓	
(Mazzon and Cavallaro, 2012)						✓	✓		

channels from the YUV space are used in (Jeong and Jaynes, 2008), where a Gaussian Mixture Model is applied to find the most relevant color clusters, whose centers are adopted as descriptors. The Dominant Color Descriptor (DCD) (Bak et al., 2010) and the Major Color Spectrum Histogram Representation (MCSHR) (Madden et al., 2007) compute the most recurrent RGB color values that are then used to represent a patch. Moreover, Maximally Stable Color Regions (MSCR) (Farenzena et al., 2010) extracts the homogeneous color in the person patch by grouping neighboring color blobs. Finally, camera parameters and reflectance of the objects' surface can be studied to obtain the main appearance characteristic of the target (Javed et al., 2008). DCD, MCSHR, MSCR and object reflectance are features applicable only when a person is captured at medium/high resolution (i.e. larger than 100×40 pixels) and there is a full body visibility.

The spatial distribution of the intensities in a person patch can be a key feature for person re-identification. Gabor and Schmid filters (Fig. 3) define two kernels for *texture* extraction applied to the luminance channel (Prosser et al., 2010; Zheng et al., 2011; Gray and Tao, 2008). Gabor filters are linear filters similar to the way the human visual system is expected to describe horizontal and vertical structures, while Schmid filters are rotational invariant Gabor-like filters. HAAR-like features can be used to extract relevant textural information from the person patch with the aim to find recurrent color distributions (Bak et al., 2010). Furthermore, the *ratios* between different regions in a patch can be used as a discriminative feature. Ratios of colors, ratios of oriented gradients and ratios of saliency maps can also be used as textural features (Berdugo et al., 2010). Similarly, Recurrent High-Structured Patches (RHSP) extract the most common blobs in the person patch (Farenzena et al., 2010); in addition to this, salient spatio-temporal edges (edgels) obtained from watershed segmentation carry information of the dominant boundary and of ratios between RGB channels (Gheissari et al., 2006). The distribution of spatial patches can be directly extracted in the frequency domain where Discrete Cosine Transform (DCT) coefficients can be used as textural features (Bauml et al., 2010). Finally, spatial patch distribution can be extracted by computing the first and the second derivatives of the person patch resulting in a covariance matrix (Wang et al., 2007; Kuo et al., 2010). Symmetric regions of a patch can be an alternative to the covariance matrix. The symmetry within the person patch can also be exploited in the extraction of local features, weighting each feature based on their position with respect to the symmetric part (Farenzena et al., 2010). These filtering methods are robust to illumination changes but cannot deal with large pose changes. In particular, Gabor and Schmid filters, and HAAR-like features are local descriptors suitable for small patches, while the ratios, RHSP, salient edgels, DCT coefficients, and covariance matrix can only be applied to people patches at medium/high resolution. Furthermore, a Histogram of Oriented Gradients (HOG) gives information on the orientation of the edges in the patch (Wang et al., 2007; Kuo et al., 2010), creating a feature that models the shape of the object by its edge distribution. However, HOG features are only invariant to changes in illumination and not to changes in pose and scale. The *silhouette* of a person has also been used when cameras are geometrically calibrated.

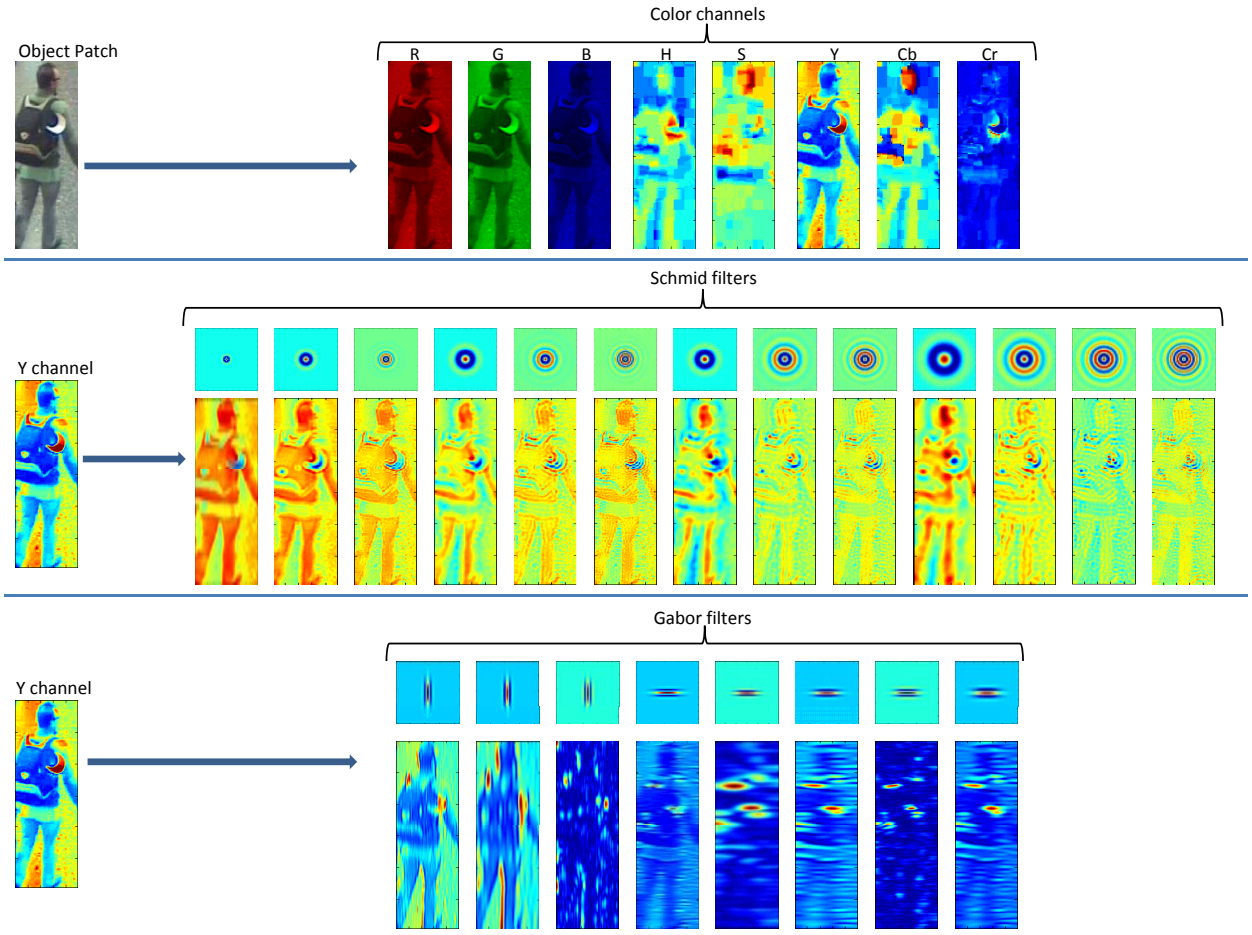


Figure 3: Example of color and texture features extraction. Color features can be extracted from different channels (top row). Textural features can be extracted by applying Schmid (middle row) and Gabor (bottom row) filters on the Y channel (Gray and Tao, 2008).

The bounding box around each person coming from single-camera tracking can be exploited by extracting the angle formed by the vertical edge and the diagonal of the bounding box (Cheng et al., 2009). A more general feature is the height of the target when calibration information is available (Berdugo et al., 2010). Finally, *interest points* can be used for re-identification in case of variations in scale, pose and illumination (Bauml and Stiefelhagen, 2011). Examples are SIFT (Teixeira and Corte-Real, 2009), SURF-like features (Hamdoun et al., 2008; Oliveira and Luiz, 2009) and the Hessian Affine invariant operator (Gheissari et al., 2006).

When intra-camera tracking information is available, features extracted from single images can be grouped over time either by *temporal* accumulation (Hamdoun et al., 2008) or by clustering (Farenzena et al., 2010). Then, the most representative patch of the group is kept as representative for the specific person. A spatio-temporal over-segmentation of patches over 10 frames can be used to create a signature for each person (Gheissari et al., 2006). However, the most common approach is to keep all the available features extracted from single patches over time and then perform association by analyzing the similarity among all the available features (Javed et al., 2008; Berdugo et al., 2010; Kuo et al., 2010). Features can also be incrementally updated over time, for example using Incremental MCSHR (IMCSHR) that updates MCSHR in order to increase robustness to abrupt changes in illumination (Madden et al., 2007). Finally, features extracted from patches of the same person over time can be used as a set of positive samples for training a learning based method (Bak et al., 2010). In general, using temporal information, the effects of light variations within the same camera and short occlusion of people are reduced because more representative features for

each target are created.

2.2. Cross-camera calibration

Cross-camera calibration includes color calibration and spatio-temporal calibration. Different illumination conditions across cameras can be compensated by robust features, as discussed above, and via color calibration where cross-camera *color calibration* models the color relationship between camera pairs (Porikli, 2003). This approach requires a learning stage where, for each camera, a relationship must be found and updated over time to cope with daily changes in the lighting conditions. Examples of color calibration include the Brightness Transfer Function (BTF) (Javed et al., 2008) and the Color Transfer Function (Jeong and Jaynes, 2008). It is demonstrated that all BTFs lie in a low dimensional space that is discovered using Principal Component Analysis (PCA) on RGB color intensities (Javed et al., 2008). In this case, color calibration is based on a linear function. An improvement of this approach is the use of the Cumulative BTF (CBTF) where the contribution of less common training samples is taken into account (Prosser et al., 2008). On the contrary, clustering on the chromaticity space can be used to find an affine color calibration transformation (Jeong and Jaynes, 2008). Color calibration can perform well in case of large inter-camera illumination changes, however it can only be applied to scenes where abrupt illumination changes are unlikely to happen.

The knowledge of the environment in which the cameras are deployed can be used to restrict the re-identification task within a certain time interval and certain regions of the monitored scenario, by estimating when and where people are going to reappear in the next camera (*spatio-temporal calibration*). In (Javed et al., 2008; Kuo et al., 2010), the learnt information are the average traveling time across cameras and the expected entry/exit points in the scene, while Prosser *et al.* (Prosser et al., 2008) select entry/exit regions manually. Learning the traveling time across cameras can be complemented by the learning of probable entry/exit regions in the camera network (Javed et al., 2008). However, when the relative camera positions are known, people location and speed can be discriminative features for each person (Cheng et al., 2009). The main limitation of these approaches is that they are only suitable for scenarios where non-observed regions are easy to model and people always follow the most common paths.

2.3. Association

The core of a person re-identification method is the definition of how to match features of candidate people. In order to associate the same person across cameras, we can measure the feature (dis)similarity, use a trained classifier, or perform an optimization process.

Person association using direct measures estimate the point-to-point *dissimilarity* between feature vectors. The Euclidean distance is used for vectors representing color values (Farenzena et al., 2010), interest points, or hypotheses of location of people (Gheissari et al., 2006; Mazzon and Cavallaro, 2012). The Euclidean distance between two colors is also included in an ad-hoc similarity measure created to compare two DCD feature sets (Bak et al., 2010). Alternative measures are the sum of quadratic distances (Oliveira and Luiz, 2009) and the sum of absolute differences (Hamdoun et al., 2008). Other distance measures include the Kullback-Leibler Distance, KLD (Jeong and Jaynes, 2008; Berdugo et al., 2010), and the Bhattacharyya Distance, BD (Prosser et al., 2008; Farenzena et al., 2010). An additional measure derived from the Kolmogrov distance is introduced in (Madden et al., 2007) to compare IMCSHR features. Correlation between color histograms and HOGs of the objects is used in (Kuo et al., 2010). In these methods, the most challenging part is the selection of the best distance for the specific set of features usually performed by trial and error.

Approaches based on measuring similarity between feature sets are not robust to illumination changes unless cross-camera color calibration is performed, as discussed earlier. As an alternative, a *classifier* can be trained to learn the changes between cameras using labeled features. Support Vector Machines (SVM) can be employed with DCT features (Bauml et al., 2010) and SIFT (Teixeira and Corte-Real, 2009). An improvement is the Ensemble SVM, which reduces the computational cost of rankSVM for high-dimensional feature spaces besides converting the re-identification problem into a ranking problem (Prosser et al., 2010). Furthermore, AdaBoost is applied for person re-identification to learn weak classifiers based on different feature sets and to identify the most discriminative features (Gray and Tao, 2008). A different learning-based approach is based on Probabilistic Relative Distance Comparison (PRDC) (Zheng et al., 2011). PRDC maximizes the probability of correct matches while minimizing that of wrong matches by learning the best distance measure for the association. Unlike direct distances, these methods are less sensitive to feature selection. However, their results can be biased by the selection of the classifier parameters, thus making the methods less flexible to different scenarios.

Other approaches use *optimization*-based algorithms. The concept of belief/uncertainty assignment can be exploited and the decision for the association problem can be made on specific ad-hoc rules (Cheng et al., 2009). An alternative approach finds the maximum likelihood Probability Density Functions (PDF) of appearance and spatio-temporal features of different observations of the same object, where the final decision is made by split graph (Javed et al., 2008). Re-identification can also be performed by Hungarian algorithm using color, texture, and spatio-temporal features (Kuo et al., 2010), where the 'potentially' correct matches are selected by Multi Instance Learning (MIL) boosting on the spatio-temporal features. Finally, dynamic programming is used to find the fitting of body models across cameras (Gheissari et al., 2006). The main drawback of optimization-based approaches is that they operate in a batch mode and cannot be run on-line.

When analyzing re-identification algorithms using the ranking score assigned to each person, results on methods solely based on appearance usually achieve less than 40-50% (Prosser et al., 2010) for the first ranking position (the real re-identification score). Re-identification algorithms that operate in batch mode exploit also spatio-temporal features, achieving results usually over 90% (Javed et al., 2008) for the first ranking position in scenes with full body visibility and linear transition of people in non-observed regions. Nevertheless, methods solely based on appearance can be tested using single snapshots of people and they became very important when cameras are located far apart. In this case, spatio-temporal calibration is very challenging and spatio-temporal features become unreliable.

3. Re-identification in crowd: proposed approach

3.1. Overview

The proposed person re-identification method deals with crowded scenes and scenarios with challenging non-observed regions where spatio-temporal calibration is not straightforward. The method extracts appearance features from the upper part of the body and models movements in non-observed regions. This modeling is performed using the map of the site under surveillance where candidate positions for people reappearance are created. Then, the association phase integrates information from appearance features and candidate positions for re-identification.

Appearance features are extracted from a representation model defined as a vertical stripe around the head location, as in a typical surveillance setting the head and the upper body are the most frequently visible and recognizable part of a person (Fig. 4). The probability of occlusion and the presence of the background in this patch are reduced by the proposed shape while maintaining the most discriminative part of each person.

Candidate positions for people reappearance are obtained on the map by propagating people movements from the first FOV, through the non-observed regions, to the entry points of the FOV of the next camera. People trajectories in the first FOV are obtained in three steps. First, head positions over time are extracted. Then, feet trajectories are estimated by going from the head trajectories down to the ground (Lv et al., 2006). Finally, feet trajectories are projected from the image plane to the map. Here the speed of each person is estimated and we assume that this speed is maintained after exiting the first FOV (Mazzon and Cavallaro, 2012). People trajectories are then propagated toward regions of interest (landmarks) in the non-observed areas (Fig. 5(a)). We define *crossing landmarks* the regions through which people transit and *entry landmarks* the regions where people may enter the FOV of the next camera (Fig. 5(b)). These landmarks are extracted from the map of the site. We refer to the proposed motion propagation based on regions of interest as Landmark-Based Model (LBM).

3.2. The Landmark-Based Model (LBM)

Let C_1 and C_2 be two cameras with non-overlapping fields-of-view and $\mathcal{M}$ be a rough map of the monitored site (top view). Let $\mathbf{x} = (x, y) \in \mathcal{M}$ be a point on the top view, which is made up by the union of the projection of the cameras' FOVs on the ground plane (Hartley and Zisserman, 2004) and the non-observed regions.

Let people $P_1, P_2, \ldots, P_N$, where N is the number of people, move onto $\mathcal{M}$, exiting C_1 and entering C_2. If $\overline{\mathbf{x}}_i$ and $\bar{t}_i$ are the last position and time instant when person P_i was visible in C_1, and $\widehat{\mathbf{x}}_i$ and $\widehat{t}_i$ are the position and time of reappearance in C_2, our goal is to model the possible paths of P_i going from $\overline{\mathbf{x}}_i$ to $\widehat{\mathbf{x}}_i$ through the landmarks (see Sec. 3.1). Notice that $\overline{\mathbf{x}}_i$ and $\widehat{\mathbf{x}}_i$ change for every person, while the landmarks are fixed for a specific map. Also, in absence of other prior information, we assume that each landmark is equally likely to be traversed in the propagation. When person P_i exits C_1, the movement is propagated toward a first landmark, and then toward crossing and entry landmarks according to specific transition rules. These rules define how people can move through the

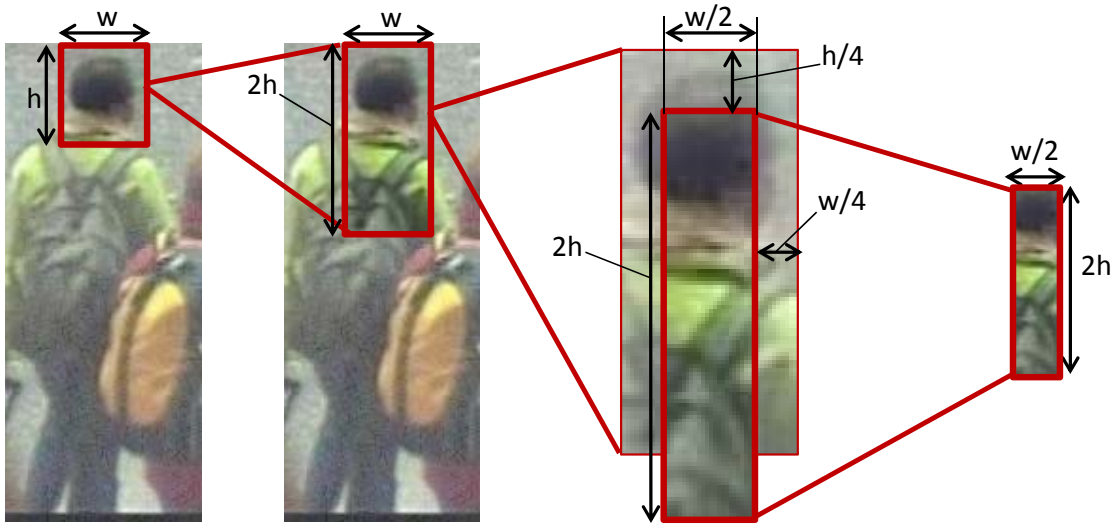


Figure 4: Spatial support for person representation. (a) Head detection bounding box. (b) Selected stripe whose height is twice the height h and half the width w of the bounding box. (c) The stripe is shifted downward by h/4 to reduce the likehood of presence of background pixels in the features used for association.

crossing landmarks and which entry landmarks can be reached from a crossing landmark (Fig. 5(c)). The entry landmarks reached after the transitions are the candidate areas for reappearance of P_i in C_2. A time step is associated to each reached entry landmark and calculated by the speed equation using the speed of people registered in the first observed region and the distance covered by the propagation through the landmarks on the top view $\mathcal{M}$.

We model crossing and entry landmarks with a set of vertices V of an oriented graph $G = (V, E)$, where E is the set of oriented edges that connect the vertices and correspond to the transitions across landmarks in $\mathcal{M}$ (Fig. 5(c)). Let $l(e)$, with $e \in E$, be the length of e. Let $A_V = \{a_1, a_2, \ldots, a_{|A_V|}\}$ with $A_V \subseteq V$ and $|A_V| > 0$, be the set of crossing landmarks and $B_V = \{b_1, b_2, \ldots, b_{|B_V|}\}$ with $B_V \subseteq V$ and $|B_V| > 0$, be the set of entry landmarks. Let $F_V = \{f_1, f_2, \ldots, f_{|F_V|}\}$ with $F_V \subseteq A_V$ and $|F_V| > 0$, be the set of vertices where the propagation of people movements can start from. Let us define

$$E_i^* = E \cup \{(\overline{\mathbf{x}}_i, F_V)\}, \tag{1}$$

where $(\overline{\mathbf{x}}_i, F_V)$ corresponds to the edges connecting $\overline{\mathbf{x}}_i$ to the set of vertices in F_V, namely the connection between the last visible position of P_i in C_1 and the vertices where the propagation can start from. Then, let us define the ordered set of edges that a person can follow to go from $\overline{\mathbf{x}}_i$ to all the entry landmarks $v \in B_V$ as:

$$\phi_i^k = \left(e_1, e_2, \ldots, e_h, \ldots, e_{|\phi_i^k|}\right), \tag{2}$$

where $e_h \in E_i^*$; $k = 1, 2, \ldots, |\Phi_i|$ and Φ_i is the set of all possible paths that person P_i can follow; $e_1 = (\overline{\mathbf{x}}_i, F_V)$ is the first edge of the sequence; $e_{|\phi_i^k|} = (A_V, B_V)$ indicates that the last edge of ϕ_i^k must go toward an entry landmark; and the edges from e_2 to $e_{|\phi_i^k|-1}$ are selected according to the transition rules. We now accumulate the time needed for person P_i to travel through a possible path ϕ_i^k using the speed equation:

$$t_{\phi_i^k} = \sum_{h=1}^{|\phi_i^k|} \frac{l(e_h)}{s_i}, \tag{3}$$

where s_i is the maximum speed calculated within a time window of T_s frames in C_1. The sum of the time step when person P_i exits C_1, $\bar{t}_i$, and the time required for P_i to traverse ϕ_i^k, $t_{\phi_i^k}$, defines the time step when person P_i reaches C_2

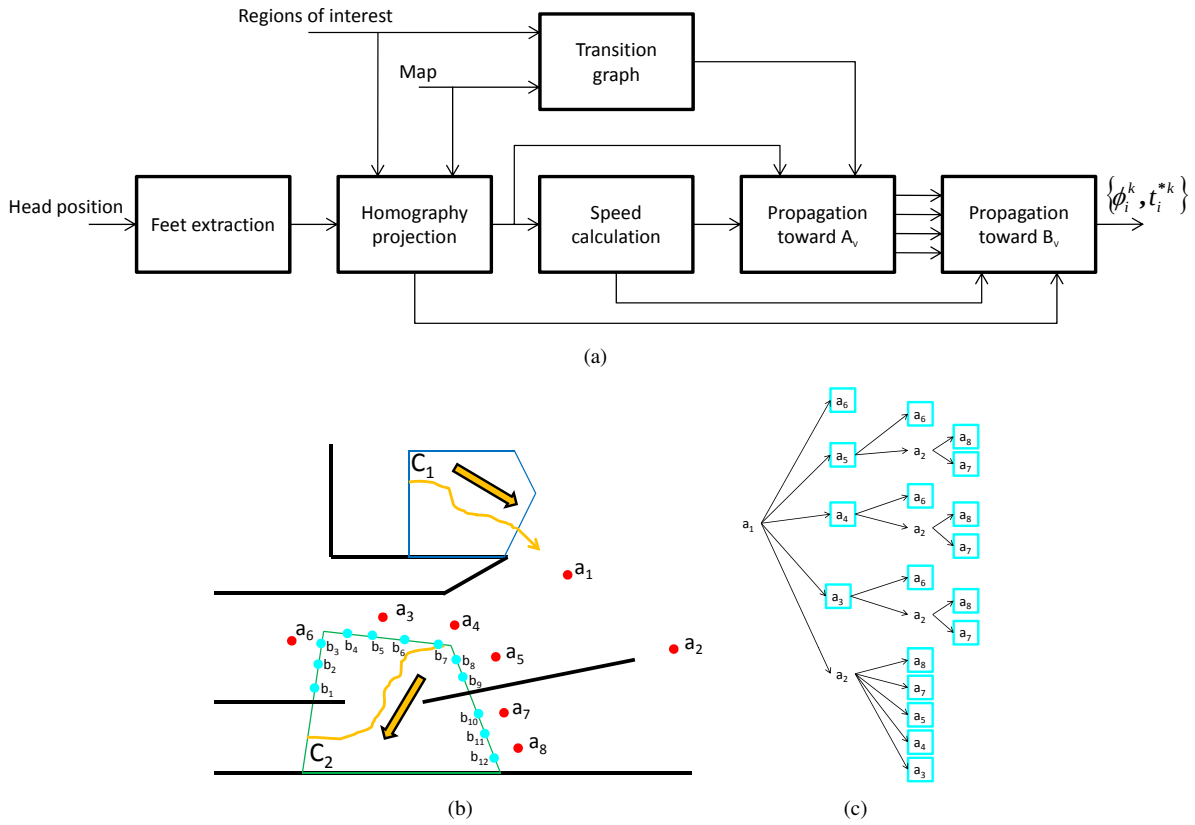


Figure 5: Landmark-Based Model (LBM): (a) Block diagram. (b) Example of setup (black line: environment map; blue and green line: FOV of Camera 1 (C_1) and Camera 2 (C_2); orange lines: trajectories; orange arrow: direction of motion; red dots: crossing landmarks; cyan dots: entry landmarks). (c) Transition graph ($\{a1, a2, \ldots, a_8\}$: crossing landmarks; cyan border: crossing landmarks from where entry landmarks can be reached).

if ϕ_i^k is traversed:

$$t_i^{*k} = \bar{t}_i + t_{\phi_i^k}. \tag{4}$$

The above process is repeated for each person exiting C_1 and going to C_2. When person P_j, with $j = 1, 2, \ldots, N$, reappears in C_2, the set of candidates for the association are the set of vertices $V_i^* = \{v_i^{*k}\} \in B_V$ reached by $e_{|\phi_i^k|}$ that satisfy

$$\widehat{t_j} - T < t_i^{*k} < \widehat{t_j} + T, \tag{5}$$

where $T \in \mathbb{N}$, thus restricting the set of possible candidates from person P_i to the closest in time to the reappearance of P_j. If T is too small, the time window would be too restrictive and the method could not account for small variations in speed. If T is too large, the time window would lose its significance to select only the "good" candidates for re-identification.

Finally, the association between P_j and the candidates from P_i is performed as a weighted sum of spatial and appearance distances. As spatial measure, we choose the Euclidean distance applied between the entering position of P_j in C_2, $\widehat{\mathbf{x}}_j$, and the candidates from P_i, v_i^{*k}. As appearance measure, we choose the Bhattacharyya distance, as it outperforms both L1-Norm and rankSVM when applied to color and texture features.

The proposed method is evaluated and compared with state-of-the-art approaches in the next section.

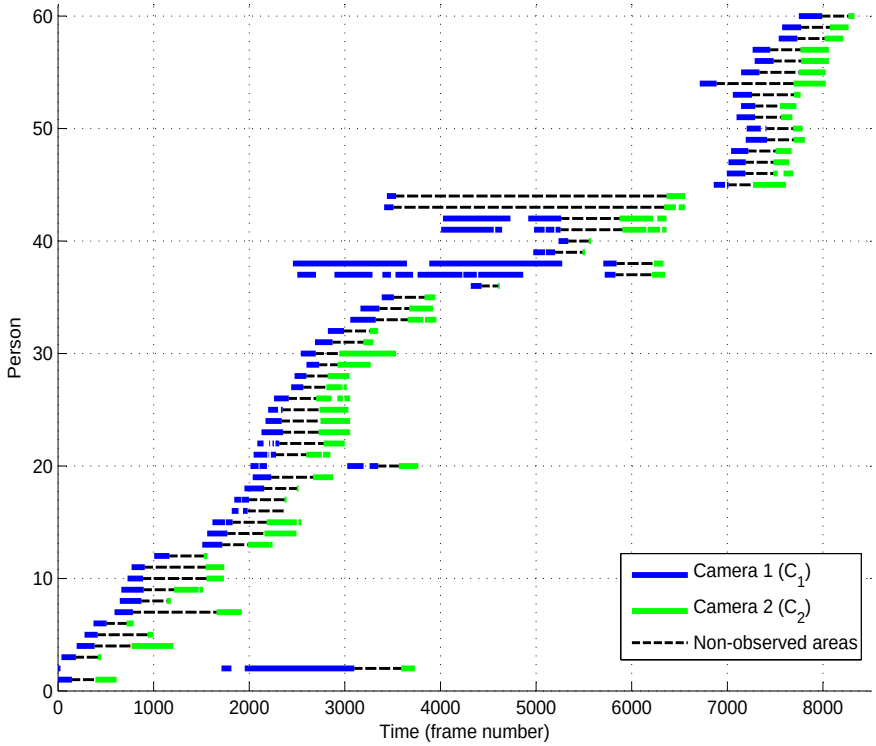


Figure 6: Time evolution of people from Camera 1 (C_1) to Camera 2 (C_2). Blue: time elapsed when a person is observed in C_1. Dotted line: time elapsed in non-observed regions. Green: time elapsed when a person is observed in C_2.

4. Experimental results

4.1. Experimental setup

In this section, we evaluate the most representative re-identification algorithms and compare them with the proposed LBM approach. Results are provided for methods based on appearance features only, spatio-temporal features only, and a combination of them. Moreover, we validate our choice to extract appearance features from a vertical stripe of the upper body (Fig. 4).

The experiments are run on the i-LIDS dataset from the London Gatwick airport (iLIDS, 2008) where similarly to previous works (Javed et al., 2008) we consider 60 people that go from C_1 to C_2. The maximum speed of people (s_i) registered in C_1 within $T_s = 50$ frames varies from 0.527 units/frame to 1.489 units/frame on $\mathcal{M}$ (with mean 0.867 units/frame and standard deviation 0.192 units/frame). Note that 1.489 units/frame corresponds to a running person. Furthermore, the two cameras present large illumination changes and people can reappear with different poses after transiting in the non-observed regions where different paths can be followed. To better understand the traveling time variability between C_1 and C_2, Fig. 6 shows the color-coded time evolution of people in the two cameras. Segments correspond to time intervals during which a person is in the FOV of a camera and is not totally occluded. It is interesting to notice that *(i)* some people stay in the FOV of C_1 for more than 1000 frames (due to the presence of shops), *(ii)* some people are visible in C_2 for only a few frames (the minimum is 4 frames), and *(iii)* the traveling time of people to go from one camera to the next is highly variable (see for example the large difference between person 36 and person 43). In addition to this, exit regions in C_1 present illumination conditions that are more similar to C_2 than the entry regions of C_1 and people are more likely to be occluded in the exit regions of C_1 than in the entry regions of C_1 due to the perspective. Hence, we divide the dataset into two sets and we perform the re-identification between entry regions of C_1 and C_2 (EN1EN2), and exit regions of C_1 and entry regions of C_2 (EX1EN2).

The appearance features for our method are extracted from the shape defined in Fig. 4: we use a concatenation

of 16-bin histograms as in (Gray and Tao, 2008; Prosser et al., 2010; Zheng et al., 2011) (see also Fig. 3). In particular, we employ 8 color channels (R,G,B,Y,Cb,Cr,H,S) from RGB, YCbCr and HSV color spaces, 8 Gabor filters (with the following parameters: $(\gamma, \theta, \lambda, \sigma^2) = (0.3, 0, 4, 2)$, $(0.3, 0, 8, 2)$, $(0.4, 0, 4, 1)$, $(0.3, \pi/2, 4, 2)$, $(0.3, \pi/2, 8, 1)$, $(0.3, \pi/2, 8, 2)$, $(0.4, \pi/2, 4, 1)$, $(0.4, \pi/2, 8, 2)$), and 13 Schmid filters (with the following parameters: $(\sigma, \tau) = (2, 1)$, $(4, 1)$, $(4, 2)$, $(6, 1)$, $(6, 2)$, $(6, 3)$, $(8, 1)$, $(8, 2)$, $(8, 3)$, $(10, 1)$, $(10, 2)$, $(10, 3)$, $(10, 4)$). For association, we compare the use of the L1-Norm, the Bhattacharyya Distance (BD), and the rankSVM (rankSVM has comparable results to the Ensemble SVM) adopted in (Prosser et al., 2008; Farenzena et al., 2010; Prosser et al., 2010; Zheng et al., 2011). The training for the rankSVM is performed with 60 people patches from C_1 and C_2 (this set does not overlap with the testing set).

4.2. Discussion

In order to evaluate our proposed representation model, we compare the results obtained with a full body model and those obtained using the shape from Fig. 4. Figures 7(a) and 7(b) show the results by Cumulative Matching Characteristic (CMC) curve. It is possible to notice that the upper body model is a more suitable shape to use for re-identification than the full body. In particular, since people are more likely to be occluded when they exit C_1, Fig. 7(b) shows a higher improvement compared to Fig. 7(a) that considers the entry of C_1. Moreover, the L1-Norm and BD show a considerable improvement from Fig. 7(a) (EN1EN2) to Fig. 7(b) (EX1EN2) compared to rankSVM that has a more stable behavior. This is because the rankSVM is a learning-based method with a cross-camera color calibration implicitly performed in the training phase and hence more robust to illumination changes.

Next, we validate the proposed LBM approach with a combination of appearance and spatial association applied to EX1EN2 where the propagation of paths is between exit regions of C_1 and entry regions of C_2. In the following experiments, we use $T = 50$ frames (corresponding to two seconds) for a fair comparison with the state-of-the-art methods (Mazzon and Cavallaro, 2012). Notice how LBM does not substantially change its performance by varying T as we obtain a mean re-identification score of 31% and standard deviation of 3% by varying T from 20 to 80 frames at steps of 5 frames.

The first comparison is between LBM with only Euclidean distance as association method, and three other methods: two spatio-temporal calibrations and one spatio-temporal modeling of people in non-observed regions. Let TTALL be the first spatio-temporal approach that calculates the average time needed for all people to go from C_1 to C_2, and consider it as the expected traveling time. Let TT4REG be the second method that divides C_2 in four entrance regions (Fig. 8) and calculates the average traveling time of people that only enter the specific region. This creates an expected traveling time for each region. To have a fair comparison, in TTALL and TT4REG we consider a correct association when a person arrives within a time interval of ± 50 frames of the expected time (the same as T). The method used for spatio-temporal modeling is the Multi-Goal Social Force Model (MGSFM) (Mazzon and Cavallaro, 2012) that applies a crowd modeling method for the trajectory propagation. People speed in MGSFM and LBM is calculated with $T_s = 25$ frames and $T_s = 50$ frames; let us call them MGSFM25 and MGSFM50, and LBM25 and LBM50, respectively. Figure 7(c) shows the CMC. Poor results are obtained with TTALL and TT4REG due to the high variability of people traveling times. In addition, LBM shows comparable results for MGSFM25 and MGSFM50 for the first 3 ranking positions and better results after ranking 3. However, LBM requires less time to be computed and a smaller number of parameters to set, thus resulting in a better applicability of the method.

Finally, we perform re-identification using LBM50 where for association, L1-Norm, BD and rankSVM are used as appearance method integrated with the Euclidean distance. Compared to only using the appearance methods, Fig 7(d) shows that LBM improves the re-identification score by 28% for L1-Norm, 28% for BD and 20% for rankSVM. These results highlight how LBM creates good candidates for re-identification. Then we weight 50% the association ranking given by appearance and 50% the one given by Euclidean distance, since it has the highest re-identification score among the possible weights (Fig. 9(a)). Fig 9(b) shows the CMC. The combination with L1-Norm gives 43%, BD 50%, and rankSVM 38% for the re-identification score that on average improves the result by 6% as compared to using only appearance and by 15% compared to using only Euclidean distance as association after performing LBM.

5. Conclusions

We proposed and validated a person re-identification method based on modeling people movements in non-observed regions using a site map and regions of interest where people are likely to transit. The two main novel con-

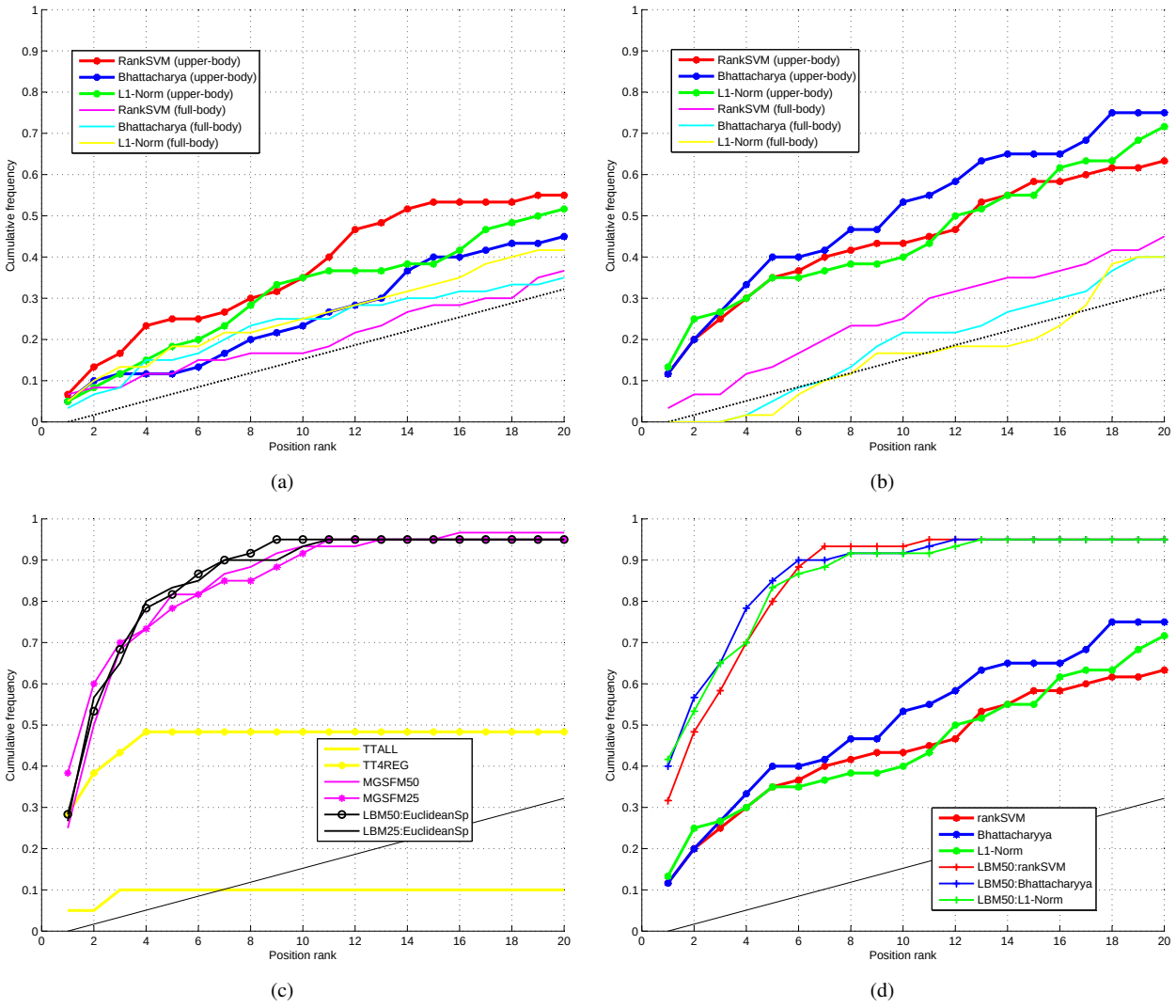


Figure 7: Cumulative Matching Characteristic (CMC) curves for person re-identification using different spatial supports for feature calculation (a)-(b), different spatio-temporal modelings of people in non-observed regions (c), and different appearance methods with and without LBM (d). Dataset: 60 people going from Camera 1 to Camera 2 in the i-LIDS dataset (iLIDS, 2008). (a)-(b): Full body is the left-most image in Fig. 4, upper body is the right-most image in Fig. 4. (a) People entering Camera 1 associated with people entering Camera 2. (b) People exiting Camera 1 associated with people entering Camera 2. (c) Multi-Goal Social Force Model (MGSFM) and LBM where association is performed using Euclidean distance; see text for details of TTALL and TT4REG. (d) Appearance methods and LBM where association is performed using different appearance methods.

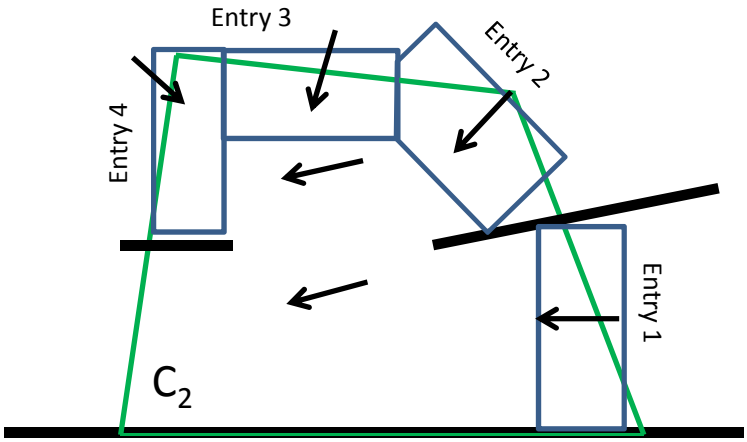
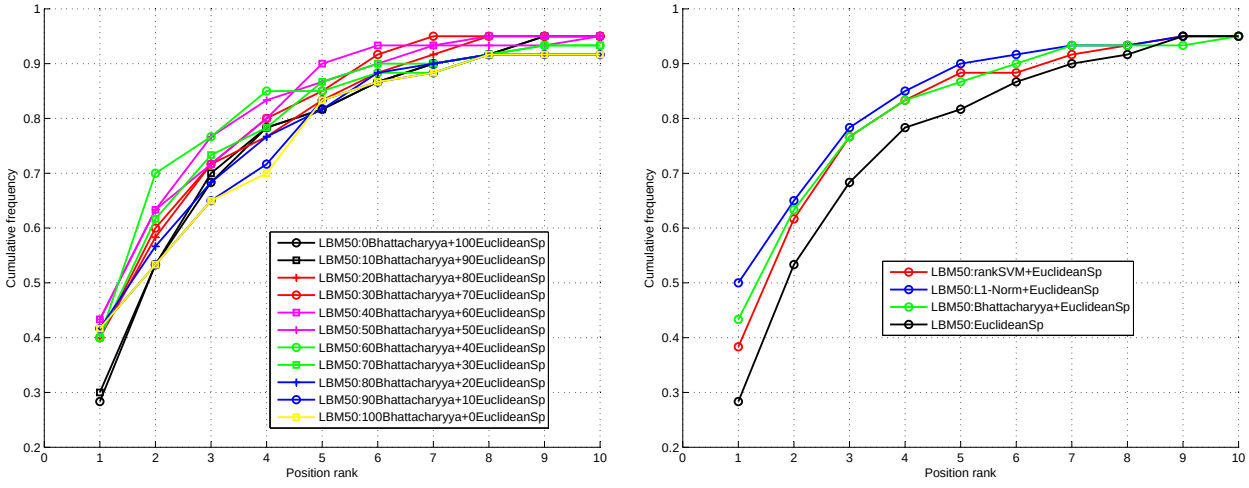


Figure 8: Entry regions of Camera 2 (C_2) (iLIDS, 2008) for spatio/temporal calibration method TT4REG (see text for details). Black lines: barriers or walls. Green line: FOV of C_2. Blue areas: possible entries. Black arrows: possible people movements.



(a) (b)

Figure 9: Cumulative Matching Characteristic (CMC) curves for person re-identification using LBM. (a) Association performed by different weighted sums of Bhattacharyya distance and Euclidean distance. (b) Association performed by weighted sum (50%-50%) of Euclidean distance and different appearance methods (compared with 100% Euclidean distance).

tributions of the proposed method are the integration of appearance information for association with spatio-temporal modeling of people movements in non-observed regions, and the extraction of features from a vertical stripe partially covering the upper body and the head, thus making the method suitable for crowded scenes. Moreover, using the proposed Landmark-Based Model (LBM), the estimation of people movements is simplified compared to the parametric crowd modeling proposed in (Mazzon and Cavallaro, 2012), where people movements are constrained by the presence of obstacles. We also organized and critically discussed the state-of-the-art for person re-identification based on features, their processing and the approaches used for association. We used a challenging dataset from the London Gatwick airport in order to compare the most relevant methods in a realistic crowded scene. In general, methods solely based on appearance features extracted on full body have performances close to random, nevertheless when they are extracted on part of the upper body they can reach 40% in the first 10 ranking positions. Using LBM, the re-identification score (rank 1) can reach 41.67% when only appearance is used, 28.33% when only motion is used, and 50% when their equally weighted sum is used for association across cameras.

As future work, we will study how different appearance features affect the performance of the proposed method, and we will extend our analysis to different scenarios, on a larger camera network and on scalability issues. We will use a dataset with people that only appear in one camera, creating the possibility to have false positive re-identifications (Bauml and Stiefelhagen, 2011).

Acknowledgment

Fahad Tahir was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the Education, Audiovisual & Culture Executive Agency (FPA n° 2010-0012).

References

Bak, S., Corvee, E., Bremond, F., Thonnat, M., 2010. Person re-identification using haar-based and dcd-based signature, in: Workshop on Activity Monitoring by Multi-camera Surveillance Systems (AMMCSS) in conjunction with IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, USA.

Bauml, M., Bernardin, K., Fischer, M., Ekenel, H.K., 2010. Multi-pose face recognition for person retrieval in camera networks, in: IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, USA.

Bauml, M., Stiefelhagen, R., 2011. Evaluation of local features for person re-identification in image sequences, in: IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), Klagenfurt, Austria.

Berdugo, G., Soceanu, O., Moshe, Y., Rudoy, D., Dvir, I., 2010. Object re-identification in real world scenarios across multiple non-overlapping cameras, in: European Signal Processing Conf. (EUSIPCO), Aalborg, Denmark.

Cheng, Y., Zhou, W., Wang, Y., Zhao, C., Zhang, S., 2009. Multi-camera-based object handoff using decision-level fusion, in: Int. congress on Image and Signal Processing (CISP), Tianjin, China.

Doretto, G., Sebastian, T., Tu, P., Rittscher, J., 2011. Appearance-based person reidentification in camera networks: problem overview and current approaches. Journal of Ambient Intelligence and Humanized Computing 2(2), 127 – 151.

Enzweiler, M., Gavrila, D.M., 2009. Monocular pedestrian detection: Survey and experiments. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(12), 2179 – 2195.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010. Person re-identification by symmetry-driven accumulation of local features, in: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Francisco, CA, USA.

Gheissari, N., Sebastian, T.B., Tu, P.H., Rittscher, J., Hartley, R., 2006. Person reidentification using spatiotemporal appearance, in: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, New York, NY, USA.

Gray, D., Tao, H., 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proc. of Europ. Conf. on Computer Vision, Marseille, France.

Hamdoun, O., Moutarde, F., Stanciulescu, B., 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 Int. Conf. on Distributed Smart Cameras (ICDSC), Stanford, CA, USA.

Hartley, R., Zisserman, A., 2004. Multiple view geometry in computer vision. Second ed. Cambridge University Press (UK).

iLIDS, 2008. Home Office multiple camera tracking scenario definition (UK).

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

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

Kuo, C.H., Huang, C., Nevatia, R., 2010. Inter-camera association of multi-target tracks by on-line learned appearance affinity models, in: Proc. of Europ. Conf. on Computer Vision, Hersonissos, Crete, Greece.

Lv, F., Zhao, T., Nevatia, R., 2006. Camera calibration from video of a walking human. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(9), 1513 – 1518.

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

Mazzon, R., Cavallaro, A., 2012. Multi-camera tracking using a multi-goal social force model. Neurocomputing - Special Issue on Behaviours in Video , To appear.

Oliveira, I., Luiz, J., 2009. People re-identification in a camera network, in: IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, Chengdu, China.

Porikli, F., 2003. Inter-camera color calibration using cross-correlation model function, in: Proc. of IEEE Int. Conf. on Image Processing, Barcelona, Spain.

Prosser, B., Gong, S., Xiang, T., 2008. Multi-camera matching using bi-directional cumulative brightness transfer functions, in: Proc. of the British Machine Vision Conf., Leeds, UK.

Prosser, B., Zheng, W.S., Gong, S., Xiang, T., 2010. Person re-identification by support vector ranking, in: Proc. of the British Machine Vision Conf., Aberystwyth, UK.

Teixeira, L.F., Corte-Real, L., 2009. Video object matching across multiple independent views using local descriptors and adaptive learning. Pattern Recognition Letters 320(2), 157–167.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P., 2007. Shape and appearance context modeling, in: Proc. of IEEE Int. Conf. on Computer Vision, Rio de Janeiro, Brazil.

Zheng, W.S., Gong, S., Xiang, T., 2011. Person re-identification by probabilistic relative distance comparison, in: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA.