Retrieval System for Person Re-identification

Sławomir Bąk[1], François Brémond[1], Vasanth Bathrinarayanan[1],
Alessandro Capra[2], Davide Giacalone[2], Giuseppe Messina[2] and Antonio Buemi[2]
[1]INRIA, STARS Group, 2004 route des Lucioles, BP93, 06902 Sophia Antipolis, France
[2]STMicroelectronics, Strada Primosole 50, 95121 Catania, Italy

Keywords: Re-identification, Retrieval, Detection, Covariance Descriptor, Brownian Descriptor.

Abstract: This paper addresses the problem of person re-identification and its application to a real world scenario. We introduce a retrieval system that helps a human operator in browsing a video content. This system is designed for determining whether a given person of interest has already appeared over a network of cameras. In contrast to most of state of the art approaches we do not focus on searching the best strategy for feature matching between camera pairs, but we explore techniques that can perform relatively well in a whole network of cameras. This work is devoted to analyze current state of the art algorithms and technologies, which are currently available on the market. We examine whether current techniques may help a human operator in solving this challenging task. We evaluate our system on the publicly available dataset and demonstrate practical advantages of the proposed solutions.

1 INTRODUCTION

Person re-identification (also known as multi-camera tracking) is defined as a process of determining whether a given individual has already appeared over a network of cameras (see fig. 1). This task can be considered on different levels depending on information cues, which are currently available in video analytics systems. For instance, biometrics such as face, iris or gait can be used to identify people. However, in most video surveillance scenarios such detailed information is not available due to video low-resolution or difficult segmentation (crowded environments, such as airports and metro stations). Therefore a robust modeling of a global appearance of an individual (clothing) is necessary for re-identification. This problem is particularly hard due to significant appearance changes caused by variations in view angle, lighting conditions and different person pose.

Owing to this complexity, current state of the art approaches have relatively low retrieval accuracy, thus a fully automated system is still unattainable. However, we propose a retrieval tool that helps a human operator to solve the re-identification task (see section 3). In this paper we discuss different techniques for automatic detection and tracking, while comparing their performance with the ground truth data. We propose a new 3D bar charts for evaluating and displaying recognition results for a large network of cameras (section 4).

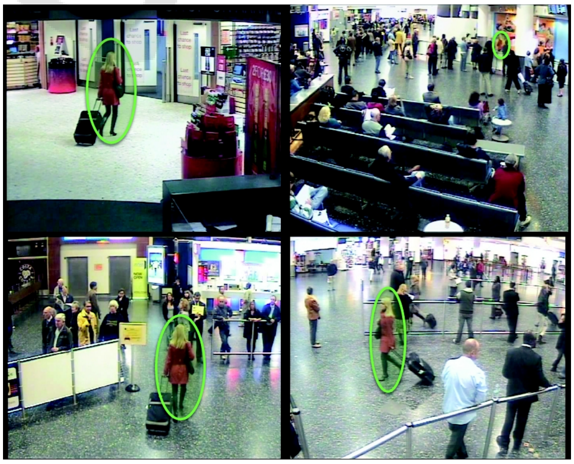


Figure 1: Person re-identification task: the system should be able to match appearances of a person of interest extracted from non-overlapping cameras.

2 RELATED WORK

Person re-identification approaches concentrate either on *metric learning* regardless of the representation choice (Dikmen et al., 2010; Zheng et al., 2011;

Bąk S., Brémond F., Bathrinarayanan V., Capra A., Giacalone D., Messina G. and Buemi A..
Retrieval System for Person Re-identification.
DOI: 10.5220/0004871807320738
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (PANORAMA-2014), pages 732-738
ISBN: 978-989-758-004-8

Koestinger et al., 2012), or on *feature modeling*, while producing a distinctive and invariant representation for appearance matching (Bak et al., 2011b; Bazzani et al., 2010; Farenzena et al., 2010). Metric learning approaches use training data to search for strategies that combine given features maximizing inter-class variation whilst minimizing intra-class variation. Instead, feature-oriented approaches concentrate on an invariant representation that should handle view point and camera changes.

Further classification of appearance-based techniques distinguishes the *single-shot* and the *multiple-shot* approaches. The former class extracts appearance using a single image (Park et al., 2006; Wang et al., 2007; Gray and Tao, 2008), while the latter employs multiple images of the same object to obtain a robust representation (Zheng et al., 2011; Bazzani et al., 2010; Farenzena et al., 2010; Gheissari et al., 2006; Prosser et al., 2010).

Unfortunately, *metric learning* approaches need training data (hundreds of image pairs with the same individual registered by different cameras) that might be difficult to acquire in a real world scenario . Moreover, these approaches focus on learning a function that transfers features space from the first camera to the second one, introducing requirement of training $\binom{c}{2} = \frac{c!}{2!(c-2)!}$ transfer functions for c cameras. This makes these solutions difficult to apply in video analytics systems.

Designing a retrieval system for a network of cameras, we believe that *multiple-shot* approaches are better choice than *single-shot*. Multiple-shot approaches take advantage of several images (video sequence) and can provide more reliable description of a target, reflecting video surveillance scenarios.

***Multiple-shot* Approaches.** In (Hirzer et al., 2011), every individual is represented by two models: descriptive and discriminative. The discriminative model is learned using the descriptive model as an assistance. In (Gheissari et al., 2006), a spatiotemporal graph is generated for ten consecutive frames to group spatiotemporally similar regions. Then, a clustering method is applied to capture the local descriptions over time and to improve matching accuracy. In (Farenzena et al., 2010), the authors propose the feature-oriented approach, which combines three features: (1) chromatic content (HSV histogram); (2) maximally stable color regions (MSCR) and (3) recurrent highly structured patches (RHSP). The extracted features are weighted using the idea that features closer to the bodies' axes of symmetry are more robust against scene clutter. Recurrent patches are presented in (Bazzani et al., 2010). Using epitome analysis, highly informative patches are extracted from a set of images. In (Cheng et al., 2011), the authors show that features are not as important as precise body parts detection, looking for part-to-part correspondences. Finally, in (Bak et al., 2011b; Bak et al., 2012) we can find re-identification algorithms based on covariance descriptor. The performance of the covariance descriptor is found to be superior to other methods, as rotation and illumination changes are absorbed by the covariance matrix. In our retrieval framework we employ the similar model to (Bak et al., 2011b), while evaluating different kind of image descriptors (details can be found in section 4).

3 RETRIEVAL SYSTEM

Our proposed system (fig. 2) allows a human operator to browse images of people extracted from a network of cameras: to detect a person on one camera and to re-detect the same person few minutes later on another camera. The main stream is displayed on the left of the screen, while retrieval results are shown on the right. The results show lists of the most similar signatures extracted from each camera (green boxes indicate the correctly retrieved person). Below the main stream window a topology of the camera network is displayed. Detection and single camera tracking (see the main stream) are fully automatic. The human operator only needs to select a person of interest, thus producing retrieval results (right screen). The operator can easily see a preview of the retrieval results and can go directly to the original video content.

The retrieval engine is based on state of the art appearance models that provide an object representation, called *signature*. This signature should be invariant to camera changes, while extracting the discrminative characteristics of a person of interest (Bak et al., 2011b; Bak et al., 2012). However, before extracting signature, we should be able to automatically determine whether an image contains people. This task is called *person detection* and it plays a very important role. The results of this step have a significant influence on the recognition/retrieval algorithms. In our framework, we evaluate two detectors. DPM, which is a popular state of the art detector and the algorithm developed by ST Microelectonics, which is mostly based on motion segmentation.

3.1 Person Detection

Person detection is considered among the hardest examples of object detection problems. The articulated structure and variable appearance of the human body,

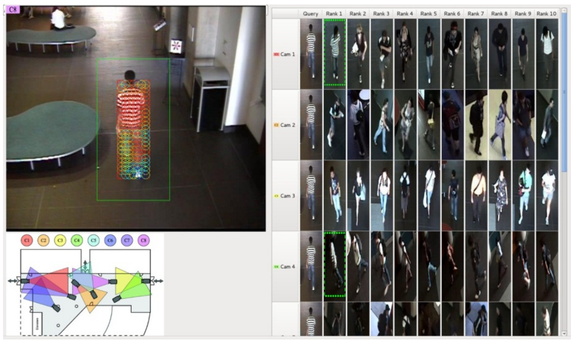


Figure 2: Person retrieval system.

combined with illumination and pose variations, contribute to the complexity of the problem. Person detection algorithm is critical in the retrieval framework, as the quality of detection has direct impact on the accuracy of retrieval results. For evaluating our framework, we employed the following detectors.

DPM is a state of the art object detector referred to as *discriminatively trained deformable part models* (Felzenszwalb et al., 2010; Girshick et al.,). This detector consists of disriminative sliding-window classifiers, predicting location of different parts of the object (*e.g.* for humans we have previously learned body part detectors). The strength of DPM comes from its ability to search through exponential number of different part-configurations, and finding the optimal one in a very efficient way. Although DPM provides state of the art performance on many evaluation datasets, its time-complexity makes it difficult to apply in real-time systems.

STMicroelectronics has investigated a motion-based detector. The algorithm has been designed in order to be implemented in embedded systems, constrains are low complexity and low memory requirements so that processing can be performed in real time on video up to HD resolution.

The first step is the detection of movement into the scene (see fig. 3(a)). Motion detection consists of a training phase, when no motion is present, to calculate the residual global motion vector of the background, and of a test phase, when actual motion has to be detected. To detect motion current and previous frame motion curves are compared. The training phase is performed on the first N frames (typically N may vary from 2 to 30). Once the motion has been detected, it is required to insulate background (fixed content in the scene) from the foreground (area of interest). Foreground is where the algorithm will search for people. This step requires that the camera must be fixed, as it is supposed that the background of the scene does not change during the acquisition. Actually, it is supposed that, in the average, the background does not change even if small variation are contemplated and managed. Background pixels are updated by weighted averaging strategy, where the weights per pixels were learned during the training phase. Shadow pixels are removed by using a technique based on HSV color space (Cucchiara et al., 2001). The background model is updated using all frames without detected objects.

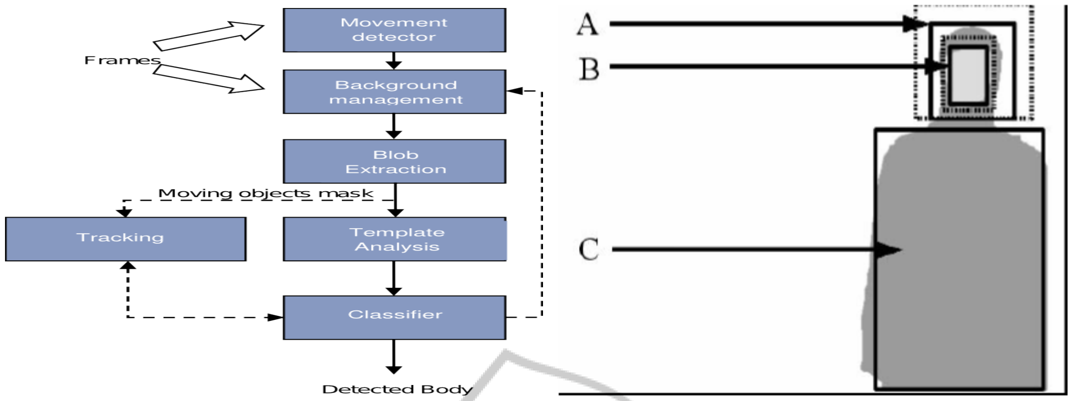
The foreground pixels are aggregated into the blob structures, on which the template matching algorithm is executed (see fig. 3(b)). The human model template module performs the task of establishing if it belongs to a human body, or not (meaning that the presence of the blob in the foreground is due to something else, like vehicle, animal, scene changes, *etc*.). For each blob a geometrical filter is applied and a test is performed in order to verify if the bounding box including the blob corresponds to a human body shape. Such test consists in comparing the bounding box, width, height and their ratio, to the following thresholds:

$$\begin{cases} h_{min} \leq h \leq h_{max}, \\ w_{min} \leq w \leq w_{max}, \\ (\frac{h}{w})_{min} \leq (\frac{h}{w}) \leq (\frac{h}{w})_{max}, \end{cases} \tag{1}$$

where h is the height and w is the width with the respective max and min thresholds. If the blob has passed the geometrical filter constrains, its shape is checked to control the human template match (Lin and S. Davis, 2010). There are different methods to perform the human template matching. For example a stylized man is shaped into the current bounding box (head + torso), and a score is updated by scrolling the bounding box: such score increases every time the blob matches the human model, pixel by pixel, according to fig. 3(b). The normalized final score is compared to a threshold and, if it is big enough, the blob is labeled as human.

3.2 Person Re-identification

After detection of a person, the system computes a human signature. Computed signatures are stored in a database, thus providing an effective interface for a human operator to search the most similar signatures to a signature of interest. Computing effective signatures that allow browsing similar people through a camera network is particularly hard due to significant appearance changes caused by variations in view angle, lighting conditions and different person pose. In this work, we follow a dense descriptors philosophy (Bak et al., 2011b; Dalal and Triggs, 2005). Every human image is scaled into a fixed size window



(a) pipeline

(b) template matching

Figure 3: STMicroelectronics person detector. (a) the main steps of the algorithm (b) the matching between the blob and a head+torso template: A: Head region with foreground pixels, and surrounding area with background pixels; B: Skin region and surrounding area; C: Foreground body region.

of 64 × 192 pixels. Then, an image is divided into a dense grid structure with overlapping spatial square sub-regions. The set of rectangular sub-regions $\mathbf{P}$ is produced by shifting 32 × 32 regions with 16 pixels step (up and down). It gives $|\mathbf{P}| = 33$ overlapping rectangular sub-regions. First, such dense representation makes the signature robust to partial occlusions. Second, as the grid structure, it contains relevant information on spatial correlations between rectangular sub-regions, which is essential to carry out discriminative power of the signature. From each sub-region, we extract 5 descriptors; three histogram-based descriptors: (1) COLOR$_{RGB}$ histogram, (2) LBP histogram (Wang et al., 2009) and (3) HOG histogram (Dalal and Triggs, 2005), and two correlation-based descriptors using (Bak et al., 2011b) feature maps: (4) COVARIANCE (Tuzel et al., 2006) and (5) BROWNIAN (Bak et al., 2014). Descriptor values are simply averaged while using several subject images (COVARIANCE is averaged on a Riemannian manifold). This provides us 5 different types of signatures, which are evaluated in the following section.

4 EXPERIMENTAL RESULTS

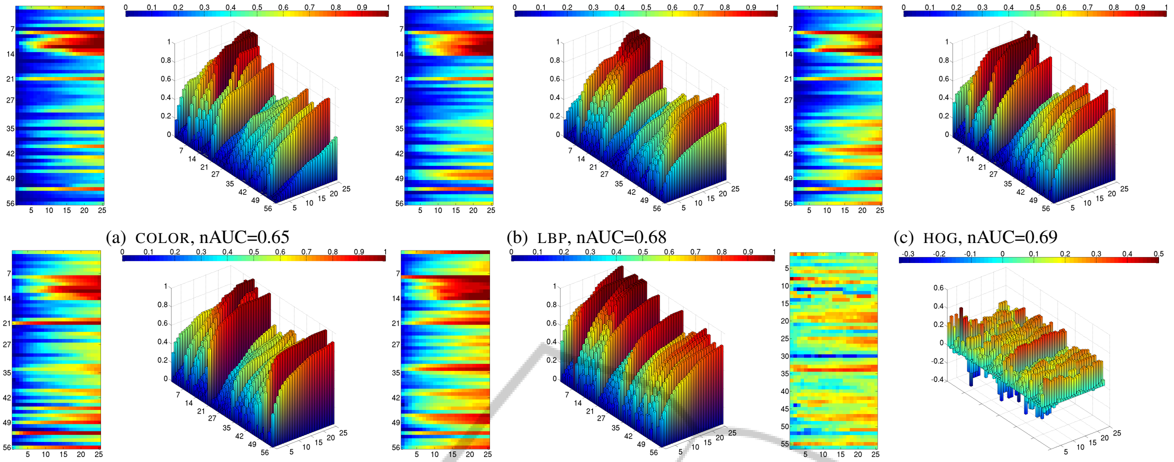
This section focuses on two tasks:

- we evaluate different types of signatures (based on different image descriptors) for a human retrieval,
- we analyze the performance drop in comparison to ground truth data, while employing automatic person detectors: DPM and STMicroelectronics detector (see section 3.1).

4.1 Re-identification Data

During the past few years person re-identification has been the focus of intense research bringing new metrics and datasets for evaluation. The most extensively used datasets are VIPER (Gray et al., 2007), ETHZ (Ess et al., 2007), i-LIDS (Zheng et al., 2009) and i-LIDS-MA/AA (Bak et al., 2011a). Although these datasets have their merits, they consist only of few camera views (maximally two) and contain only few images per person. VIPER contains two views of 632 pedestrians but it is is limited to a single image. In ETHZ, although the video sequences are acquired from moving camera, all pedestrians are extracted from the same sensor. This significantly simplifies the task of re-identification. While i-LIDS-MA/AA are designed for evaluating multiple-shot case (contain significant number of images per individual), they still consist only of two camera views. In the result, we decided to evaluate our re-identification system on a new dataset SAIVT-SOFTBIO (Bialkowski et al., 2012) .

SOFTBIO (Bialkowski et al., 2012). This database consists of 152 people moving through a network of 8 cameras. Subjects travel in uncontrolled manner thus most of subjects appear only in a subset of the camera network. This provides a highly unconstrained environment reflecting a real-world scenario. In average, each subject is registered by 400 frames spanning up to 8 camera views in challenging surveillance conditions. Each camera captures data at 25 frames per second at resolution of 704 × 576 pixels. Although some cameras have overlap, we do not use this information while testing re-identification algorithms. Authors (Bialkowski et al., 2012) provide XML files



(a) COLOR, nAUC=0.65 (b) LBP, nAUC=0.68 (c) HOG, nAUC=0.69

(d) COVARIANCE, nAUC=0.73 (e) BROWNIAN, nAUC=0.79 (f) BROW. - COV., ΔnAUC=0.06

Figure 4: Descriptor performances as CMC bars for 56 camera pairs (a-e) of SAIVT-SOFTBIO dataset. nAUC is a weighted (by gallery size) average of nAUC obtained by each pair of cameras. For each descriptor the top view and 3D chart is presented. Red color indicates high recognition accuracy. For each descriptor we can notice the red region on the top view (see rows 7 – 14). This is the retrieval result for the second camera, in which only few subjects were registered (29 out of 152). The rest of cameras is more balanced (about 100 subjects per camera). (f) illustrates the difference between BROWNIAN and COVARIANCE. We can notice that BROWNIAN performed better for most of camera pairs (bluish color correspond to opposite case).

Table 1: Descriptor performance comparison on SAIVT-SOFTBIO dataset. Values correspond to the recognition accuracy averaged among all 56 pairs of cameras at different ranks r.

DESCRIPTOR	$r = 1$	$r = 5$	$r = 10$	$r = 25$
BROWNIAN	**15.96%**	**33.53%**	**47.37%**	**70.09%**
COLOR	6.12%	19.09%	29.79%	50.60%
LBP	8.30%	22.62%	32.92%	53.91%
HOG	10.02%	24.73%	37.64%	60.89%
COVARIANCE	12.83%	28.65%	40.09%	64.13%

with annotations given by coarse bounding boxes indicating the location of the subjects.

4.2 Evaluation Metrics

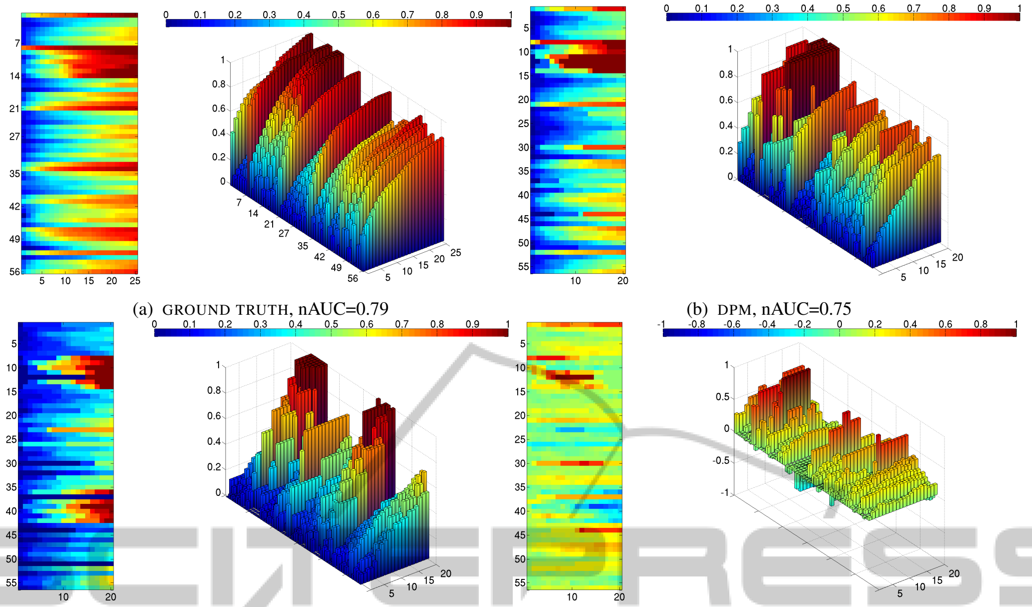
Usually, the results of re-identification are analyzed in terms of recognition rate, using the averaged *cumulative matching characteristic* (CMC) curve (Gray et al., 2007). The CMC curve represents the expectation of finding the correct match in the top n matches. Additionally, some authors also report a quantitative scalar of CMC curve obtained by the normalized area under CMC curve (nAUC). In this paper, instead of using averaged CMC curves, we display the results using 3D bar-chart (see fig. 4). The horizontal axis corresponds to recognition accuracy, while on vertical axis the first 25 ranks are presented for each camera pair (*e.g.* having 8 cameras we actually can produce 56 CMC bar series that present recognition accuracy for each camera pair). We also color the CMC bars *w.r.t.* recognition accuracy and display it as a top-view image. In the result we can see that re-identification accuracy might be strongly associated with a particular camera pair (similar/non-similar camera view, resolution, the number of registered subjects).

4.3 Image Descriptors

Figure 4 illustrates the retrieval results for different kinds of descriptors. From the results it is apparent that Brownian descriptor outperforms the rest of descriptors. In the result, in the next section we employ only this descriptor. Table 1 shows the averaged (among all 56 camera pairs) recognition accuracy *w.r.t.* to the rank. We can see that the Brownian descriptor consistently achieves the best performance for all ranks.

4.4 People Detectors

We performed retrieval task using 3 detection data: (1) annotations available in SAIVT-SOFTBIO (coarse



(a) GROUND TRUTH, nAUC=0.79

(b) DPM, nAUC=0.75

(c) STM, nAUC=0.57

(d) DPM - STM, ΔnAUC=0.18

Figure 5: Comparison of person detectors using BROWNIAN descriptor.

bounding boxes indicating the location of the subjects have been annotated every 20 frames and intermediate frames locations were interpolated); (2) DPM results (we only took into account the body bounding boxes) and (3) output provided by STMicroelectronics. The signatures were extracted using the dense grid of BROWNIAN covariances. Results are presented in fig. 5. We can notice that DPM performs relatively well to annotated data. This is a very promising result. Although DPM is difficult to apply in real-time tracking systems due to its time-complexity, the re-identification task does not directly require real-time performance (signatures could be computed in an off-line mode), thus this performance could be achieved by the automatic system.

Using STMicroelectronics algorithm we obtain worse performance. The main reason of such result is due to background images provided by the dataset. For several video sequences, there is a significant difference between a background image and a corresponding video sequence (the background image is not updated, thus motion detection algorithm returns noisy blobs due to *e.g.* new objects (belonging to the background) on the scene and significant illumination changes). Moreover some of video sequences contain gaps of several frames what has substantial impact on the motion detection algorithm. As STMicroelectronics algorithm is purely based on motion, the above mentioned issues might cause noisy detection results, decreasing recognition accuracy. We believe that correct (updated) background images and full video streams (no gaps in video sequences) could significantly improve the detection and the re-identification quality.

5 CONCLUSIONS

We described our person re-identification system, while evaluating (1) different kinds of descriptors for representing human signatures and (2) different people detectors. The results are illustrated using 3D bar-charts that allow to display recognition accuracy *w.r.t.* camera pairs. Our framework was evaluated on challenging SAIVT-SOFTBIO dataset, which provides the unconstrained environment reflecting a real-world scenario. We obtained promising results, while testing state of the art algorithms. The best performance for the fully automatic system was achieved by combining the DPM detector with BROWNIAN descriptor. In the future, we plan to combine the notion of motion with DPM detections. This would allow to extract only the features, which surely belong to foreground regions.

ACKNOWLEDGEMENTS

This work has been supported by PANORAMA and CENTAUR European projects.

REFERENCES

Bak, S., Charpiat, G., Corvee, E., Bremond, F., and Thonnat, M. (2012). Learning to match appearances by correlations in a covariance metric space. In *ECCV*.

Bak, S., Corvee, E., Bremond, F., and Thonnat, M. (2011a). Boosted human re-identification using riemannian manifolds. *Image and Vision Computing*.

Bak, S., Corvee, E., Bremond, F., and Thonnat, M. (2011b). Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*.

Bak, S., Kumar, F. R., and Bremond, F. (2014). Brownian descriptor: a Rich Meta-Feature for Appearance Matching. In *WACV*.

Bazzani, L., Cristani, M., Perina, A., Farenzena, M., and Murino, V. (2010). Multiple-shot person re-identification by hpe signature. In *ICPR*, pages 1413–1416.

Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., and Lucey, P. (2012). A database for person re-identification in multi-camera surveillance networks. In *DICTA*, pages 1–8.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *BMVC*, pages 68.1–68.11.

Cucchiara, R., Grana, C., Neri, G., Piccardi, M., and Prati, A. (2001). The sakbot system for moving object detection and tracking. In *Video-Based Surveillance Systems-Computer Vision and Distributed Processing*, pages 145–157.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, volume 1.

Dikmen, M., Akbas, E., Huang, T. S., and Ahuja, N. (2010). Pedestrian recognition with a learned metric. In *ACCV*, pages 501–512.

Ess, A., Leibe, B., and Van Gool, L. (2007). Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In *CVPR*, pages 1528–1535.

Girshick, R. B., Felzenszwalb, P. F., and McAllester, D. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *PETS*.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275.

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *CVPR*.

Lin, Z. and S. Davis, L. (2010). Shape-based human detection and segmentation via hierarchical part-template matching. *Transactions on Software Engineering*, pages 604–622.

Park, U., Jain, A., Kitahara, I., Kogure, K., and Hagita, N. (2006). Vise: Visual search engine using multiple networked cameras. In *ICPR*, pages 1204–1207.

Prosser, B., Zheng, W.-S., Gong, S., and Xiang, T. (2010). Person re-identification by support vector ranking. In *BMVC*, pages 21.1–21.11.

Tuzel, O., Porikli, F., and Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *ECCV*, pages 589–600.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P. (2007). Shape and appearance context modeling. In *ICCV*, pages 1–8.

Wang, X., Han, T. X., and Yan, S. (2009). An HOG-LBP human detector with partial occlusion handling. In *ICCV*.

Zheng, W.-S., Gong, S., and Xiang, T. (2009). Associating groups of people. In *BMVC*.

Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *CVPR*.