3DPeS: 3D People Dataset for Surveillance and Forensics

Davide Baltieri, Roberto Vezzani, Rita Cucchiara
D.I.I. - University of Modena and Reggio Emilia
{davide.baltieri, roberto.vezzani, rita.cucchiara}@unimore.it

ABSTRACT

The interest of the research community in creating reference datasets for performance analysis is always very high. Although new datasets, collecting large amounts of video footage are spreading in surveillance and forensics, few benchmarks with annotation data are available for testing specific tasks and especially for 3D/multi-view analysis. In this paper we present 3DPeS, a new dataset for 3D/multi-view surveillance and forensic applications. This has been designed for discussing and evaluating research results in people re-identification and other related activities (people detection, people segmentation and people tracking). The new assessed version of the dataset contains hundreds of video sequences of 200 people taken from a multi-camera distributed surveillance system over several days, with different light conditions; each person is detected multiple times and from different points of view. In surveillance scenarios, the dataset can be exploited to evaluate people reacquisition, 3D body models and people activity reconstruction algorithms. In forensics it can be adopted too, by relaxing some constraints (e.g. real time) and neglecting some information (e.g. calibration). Some results on this new dataset are presented using state of the art methods for people re-identification as a benchmark for future comparisons.

Categories and Subject Descriptors

I.4.8 [**Image Processing And Computer Vision**]: Scene Analysis —*Tracking*; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Performance, Experimentation

Keywords

Surveillance, Forensics, Dataset, Re-Identification, Reacquisition

J-HGBU'11, December 1, 2011, Scottsdale, Arizona, USA.

1. INTRODUCTION

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

For assessed research and deeply explored problems, there is the need to compare new technical solutions, vendor promises, requirements and limitations in real working conditions; typical examples are in biometrics where, although research is in continuous evolution, the market is interested in giving validation and standardization: see, as an example, the long story in the evaluation of face recognition techniques that started with the FERET [20] contest more than ten years ago. In some cases, when data are not easily available some synthetic datasets have also been proposed and largely adopted (e.g. FVC2000 [15]).

For emerging activities and open problems instead the main need is to fix some common limits to the discussion and to have an acceptable starting base to compare solutions.

Often kernel benchmark datasets are defined to stress specific algorithms such as datasets for shadow detection, pedestrian detection or other common tasks in surveillance. Among them few datasets have been proposed for 3D/multi-view re-identification in surveillance and forensics.

In this paper we propose a new dataset benchmark, namely **3DPeS** (**3D Pe**ople **S**urveillance Dataset) for surveillance and forensic in 3D environments, specifically designed for re-identification tasks, but applicable to many other tasks too, such as people detection, tracking, action analysis and trajectory analysis.

The main novelty is that data for the complete processing chain is available: the camera setting and the 3D environment reconstruction, the hundreds of recorded videos, the camera calibration parameters, the identity of the hundreds of people, detected more than one time by different points of view.

In 3DPeS also segmentation and tracking data are recorded as baseline processing, together with some soft-biometrics of the individuals (e.g. their height). In surveillance, it can be exploited to extract several performance indicators for re-identification by similar and different cameras; videos are taken at the same frame rate and with a given topologi-

cal setting so that real-time surveillance can be simulated; having 3D information, also reasoning with 3D body and environment models can be evaluated. In forensics it can be adopted by relaxing some constraints, for instance neglecting the calibration data which are normally unknown. In the next section, after a brief presentation of the state-of-the-art of datasets for re-identification, we will describe our new dataset and we will show some initial comparison results using state of the art re-identification algorithms.

2. RELATED WORKS

After many years of research, we have not yet reached the ultimate solutions for all surveillance tasks, and even less in the more recent area of video analytics for forensics. This leads to a widespread urgency in common evaluation of proposals. Both in surveillance and forensics, the problem of people *re-identification* or *consistent labeling*, i.e. the capability of associating together the views of the same person captured in different places or after temporal intervals, is still open. Among the many approaches, three different strategies could be defined, mainly depending on the camera setup and environmental conditions:

- **Biometric approaches**: the various person instances are matched together and, possibly, are assigned to the same identity by means of (hard) biometric features. Gait, faces, fingerprints, iris scans and so on are some examples adopted in real situations [11, 7]. Even if they are the most reliable and effective solutions, they require suitable sensors and a collaborative behavior of the people. Thus, in the case of common settings with surveillance cameras (low resolution, poor views, non collaborative people) they are not always applicable.

- **Geometric approaches**: when more than one camera or sensor simultaneously collects information of the same area, geometrical relations among the fields of view (e.g, homographies, epipolar lines, and so on) can be adopted to match the different detections [4, 13, 9]. If available, geometric relations guarantee strong matches or, at least, a stiff candidate selection.

- **Appearance based approaches**: in the most general case, only the appearance of the different items can be used [6, 1]. Re-identification can be correctly done only if the appearance is preserved among the views. This can be considered a soft-biometric approach (exploiting dress colors and textures, perceived heights and other similar cues). Occlusions, illumination changes, different sensor qualities, different viewpoints are some of the challenging issues which make the appearance based re-identification a hard problem.

For a comprehensive analysis of state of the art methods for people re-identification, please refer to [6, 9].

Despite this field is attracting several research groups and diverse solutions have been proposed, few datasets and metric evaluation criteria are available. Several datasets are publicly available for testing common surveillance tasks, such as pedestrian detection, object detection, scene categorization, face recognition and action and behavior analysis. The most famous are PETS [19], iLids [12], Etiseo[16], OpenVISOR[23]. The latest dataset proposed in CVPR2011 is VIRAT[17], produced by a collaborative effort of different US Labs, it contains a very large quantity of surveillance videos recorded in an unconstrained environment. Also TRECVID [18], developed mainly for multimedia analysis, covers some aspect of surveillance, like event detection. Unfortunately all these datasets are not devoted to re-identification problems. One interesting small "kernel benchmark" has been proposed in [22] with 34 peoples re-acquired by a single camera in a bus. Currently, one of the most popular and challenging dataset for people re-identification is **VIPeR**[9], which contains 632 image pairs of pedestrians taken from arbitrary viewpoints under varying illumination conditions (see Fig. 1 for some examples). The data were collected in an academic setting over the course of several months, with each image scaled down to 128x48 pixels. Due to its complexity and the low resolution only few researchers have published their results on VIPeR; actually, some matches are hard to identify even by a human, such as the third couple in Fig. 1. Currently, the best results on this dataset have been obtained by Farenzena et al [6], Gray et al [10] (the authors of the dataset) and Prosser et al [21].

Figure 1: Examples of real matches from the ViPER dataset [9].

Unfortunately, VIPeR contains two images for each target only. In surveillance scenarios instead, a sequence of frames captured by a static calibrated camera is usually available for each person instance. Integration of different views and geometric information can be exploited for the re-identification, as proposed by [8, 2] but the VIPeR dataset is not suitable to test all these approaches, or at least limits to stress their capabilities as in real settings. A small dataset was formerly defined (that could constitute a beta-version for the current benchmark), the **ViSOR** people re-identification dataset, consisting of short video clips captured with a calibrated camera[3]. To simplify the frame alignment, four frames for each clip have been manually selected, corresponding to predefined positions and postures of the people. Thus the dataset is composed by four views for each person, 200 snapshots in total. Some examples are shown in Fig. 2, where the output of the foreground segmentation is reported.

Although better suited for multi-view people reidentification, the ViSOR people re-identification dataset was still very limited: the dataset includes only 50 people, and only the preselected 4 main views (frontal, left and right side, back) are provided, silhouette segmentation is mostly very good, and all images were taken under the same light condition.

3. THE 3DPeS VIDEO BASED DATASET

The main limitation of currently annotated datasets for re-identification is that they are mainly image-based and not video based. For example, in VIPeR only two views for each person are provided, without any form of calibra-

Figure 2: Sample frames from our older re-identification dataset [3]: the blobs extracted with a motion segmentation and tracking system are shown.

tion; additionally, all images are scaled down to the same resolution, causing some distortion between body parts.

To cope with these limitations and to provide a specific dataset for all the usual steps in video surveillance (segmentation, tracking etc.) we designed a new standard dataset: **3DPeS** (**3D Pe**ople **S**urveillance Dataset). This new dataset is under development in subsequent steps, with the first step showing the more simple cases (e.g. few people at a time and little occlusions), while latter steps will add more complex scenarios. Here we will give a brief description of the current completed step (Version 1.0) and of the next step of this dataset (version 2.0) available in late 2011 on the ViSOR online repository (www.openvisor.org).

Figure 3: Sample frames from our re-identification dataset

3.1 3DPeS v1.0

The starting point of our dataset is a real surveillance setup, composed by 6 different surveillance cameras (see fig.3 and fig.4), monitoring a section of the campus of the University of Modena and Reggio Emilia. Data were collected over the course of several days. Subjects were notified of the presence of cameras, but were not coached or instructed in any way. The illumination between cameras is almost costant, but people were recorded multiple times during the course of the day, in clear light and in shadowy areas, resulting in strong variations of light conditions in some cases. The quality of the camera hardware is in line with current standards in visual surveillance, all cameras were from the same vendor and are partially calibrated (position, orientation, pixel aspect ratio and focal length are provided for each one of them). The quality of the images is mostly constant, uncompressed images with a resolution of 704x576 pixels. Depending on the camera position and orientation, people were recorded at different zoom levels.

N°of videos	500
N°of frames (average, per video)	400
N°of peoples	200
Total N°of frames	200000
Resolution	704x576
N°of video sequences (average, per person)	5
N°of cameras (average, per person)	2

Table 1: Quantitative characteristics V1.0

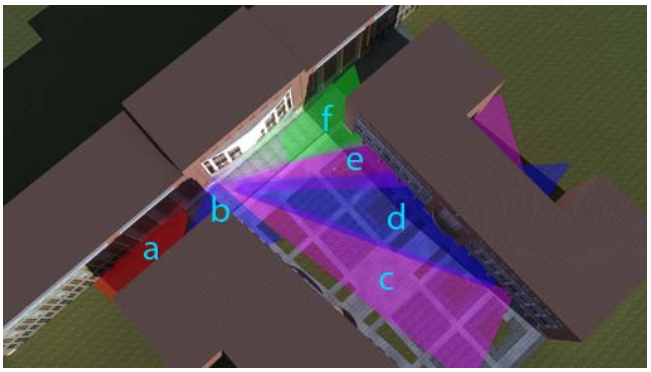


Figure 4: Camera setup

Table 1 reports some quantitative characteristics of the dataset. Annotation for the v1.0 of the dataset comprises camera parameters, person IDs and correspondences across the dataset and bounding box of the target person in the first frame of the sequence (See table 2 for additional details). In this first step each video sequence contains only the target person or a very limited number of people.

3.2 3DPeS v2.0

Our goals for the next step in the dataset development (2.0) is to reach at least 350000 frames and 300 people with full tracking annotation (trajectories, people heights etc.); Tracking annotation will be generated using the SAKBOT system [5] as a baseline. A 3D reconstruction of the surveilled area will be available in COLLADA and Lightwave format. Table 2 describe the elements of the annotation that will be present in the final version. Additionally, more challenging video sequences will be added, featuring groups of people, occlusions and different weather conditions.

4. PERFORMANCE EVALUATION

Defining common metrics and standard procedures for performance evaluation is still an open issue, despite some

Elements	V1.0	v2.0
Full Calibration for the 6 cameras	Yes	Yes
Background Model for each video	Yes	Yes
3D Space Reconstruction	Partial	Yes
Positions on Ground Plane	Partial	Yes
People Heights	Partial	Yes
Segmentation with BBox	Yes(first frame)	Yes
People Silhouettes	No	Yes(some frames)
Shadow Segmentation	No	Yes(few frames)
Trajectory	No	Yes
People IDs	Yes	Yes

Table 2: v1.0 and v2.0 annotation characteristics

tentatives done toward this direction. For detection and tracking purposes the ETISEO project [16] proposed some well accepted metrics (ETISEO was a project devoted to performance evaluation for video surveillance systems, studying the dependency between algorithms and the video characteristics). The 3DPeS Dataset can be used to evaluate the performance of every step: segmentation, tracking, shadow analysis, trajectory and action recognition and so on. For re-identification two paradigms could be adopted. Considering the re-identification as a matching problem (enhanced tracking), **precision and recall** measures could be adopted in order to evaluate how many true and false matches have been carried out. More sophisticated scores have been specifically proposed for the tracking, such as the **tracking time** and the **object ID persistence**' [16]. The first one corresponds to the percentage of time during which the reference item is detected and tracked. This metric gives us a global overview of the performances of the multi-camera tracking algorithm. Yet, it suffers from the issue that the evaluation results depend not only on the re-identification algorithms but also on the detection and single camera tracking system. The second metric qualifies the re-identification precision, evaluating how many identities have been assigned to the same real person.

The before mentioned metrics are useful to evaluate the re-identification process embedded into real time surveillance systems. For forensic activities, instead, the matching process could be done in a semi-automatic way: the user submits queries to the system looking for a set of candidates instead of a single match. Thus, the re-identification process could be considered as a ranking problem [9]. In this case, the **cumulative matching characteristic** (CMC) curve is the proper performance evaluation metric, showing how performances improve as the number of requested images increases.

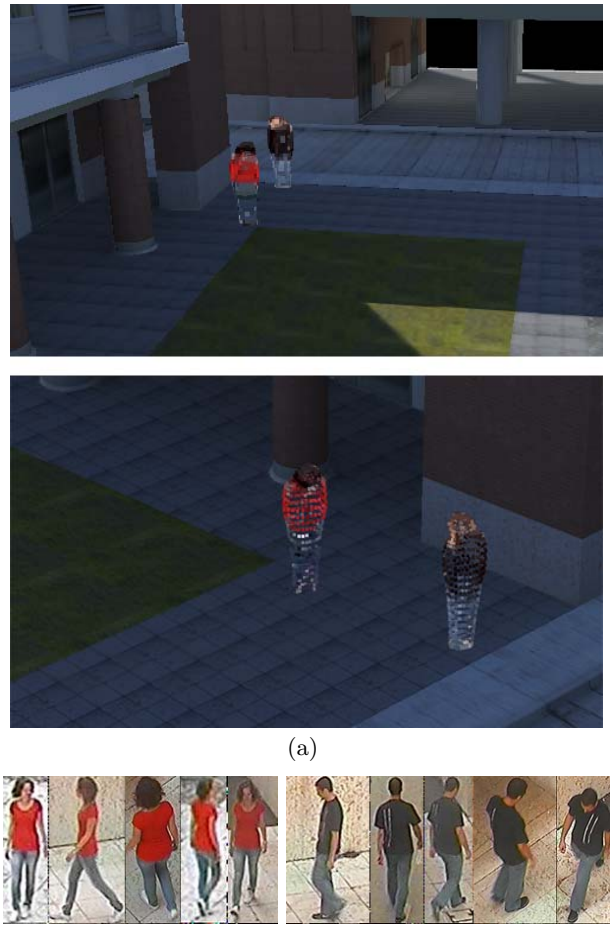
The CMC curve represents the expectation of finding the correct match in the top n matches and is analogous to the ROC curve for detection problems.

If the CMC curve for the matching function is given, the probability that any of the M best matches is correct generates the synthetic recognition rate (SRR) as follows:

$$SRR(M) = CMC(N/M) \quad (1)$$

where $CMC(k)$ is the rank k recognition rate and M is the number of pedestrians in the scene taken from a large testing dataset of size N (see fig. 6 for an example of CMC curve).

Finally, let us cite the work by Leung et al [14] about performance evaluation of re-acquisition methods specifically developed for public transport surveillance, which takes into account the a-priori knowledge of the scene and the normal people behaviors to estimate how the re-identification system can reduce the entropy of the surveillance framework.



(a)

(b)

Figure 5: (a) 3D Reconstruction of a section of the campus of the univesity of Modena and Reggio Emilia, also shown are the 3D reconstructions of the appearance of two tracked persons, using the method described in [3] (b) The frame selected by the 3D re-identification algorithm, as described in [3]

4.1 Preliminary Experiments

To show the advantages in using 3D information and multiple views for the reacquisition and re-identification of people we performed some preliminary experiments and compared two very different re-identification methods on our new dataset:

- SDALF (Symmetry Driven Accumulation of Local Features) proposed in 2010 by Farenzena et al [6].

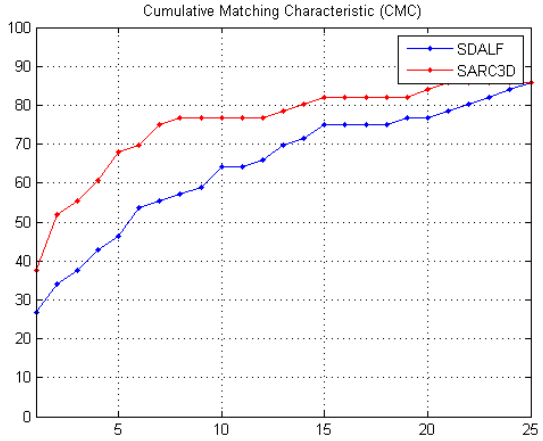


Figure 6: Results of our preliminary tests using SDALF and the SARC3D Method

- SARC3D a multi-view method based on a 3D body model we proposed in 2011 [3].

The first is a purely two dimensional method. It consists in the extraction of features that model three complementary aspects of the human appearance: the overall chromatic content (using weighted HSV histograms), the spatial arrangement of colors into stable regions (Maximally Stable Color Regions), and the presence of recurrent local motifs with high entropy. All these features are derived from different body parts, and opportunely weighted by exploiting symmetry and asymmetry perceptual principles (each appearance image is segmented into legs/torso/head using simple heuristics). We compared the aforementioned method with the multi-view 3D-based method SARC3D [3]. People are detected and tracked in each calibrated camera, with their silhouette, appearance, position and orientation extracted and used to place, scale and orientate a simplified 3D body model. For each vertex of the model a signature (color features, reliability and saliency) is computed from 2D appearance images and exploited for matching. Fig. 5(a) shows some frames of the SARC3D method in action.

In both cases the AD HOC system [24] was used to track peoples and extract their silhouette. Using the same algorithm described in [3] only few selected appearance images were automatically extracted from each video sequence and used for the re-identification: in the case of SDALF we selected randomly a single view from each video sequence, while between 3 and 5 images for each sequence were randomly selected for the creation of the Sarc3D models.

For each method we performed 10 test runs using sequences of 100 people randomly selected from the dataset. Figure 6 shows the averaged CMC curves of both methods, figure 7 shows some selected test queries. As the graph shows, by using multiple views and 3D data the reidentification performances significantly increases, especially if the number of returned ranked matches is limited. For instance at rank-1 SDALF returns 26.78% of correct matches, while SARC3D 37.51%. At rank-5 46.42% for SDALF against 67.8% for SARC3D.

Figure 7: Example queries to our re-identification database. (a) Probe image (for SARC3D this is just one of the images used for the model creation). (b) Top 10 results (sorted left to right). First row SDALF results, second row SARC3D results. The correct match is highlighted in green.

5. CONCLUSIONS

This paper presents a new dataset for 3D/multi-view surveillance and forensic analysis, especially for the evaluation of research results in people re-identification and other related activities. On this dataset, different performance analysis metrics can be exploited at frame level and at video shot level. Some comparisons can be done with different algorithms: some test are performed with the state of the art of 2D and 3D body models. These tests confirm a straightforward consideration that having the possibility to exploit more views of the same person, although with simpler or similar features, the re-identification produces better results. This consideration seems to be obvious but it must be taken into account in the new generations of surveillance and forensic platforms where, actually, video clips and thus many images can be analyzed. Of course this needs more processing, as for instance to detect the approximate people direction to cope with the alignment of the 3D model but we believe that in the future surveillance task chain this could be sufficiently assessed too, and for this reason 3D datasets will be more and more required.

6. ACKNOWLEDGMENTS

This work has been done within the THIS project with the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security.

7. REFERENCES

[1] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, 2010.

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. 3d body model construction and matching for real time people re-identification. In *Proc. of EG-IT 2010*, pages 65–71, 2010.

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identi cation. In *Proc. of ICIAP*, Ravenna, Italy, Sept. 2011.

[4] S. Calderara, A. Prati, and R. Cucchiara. HECOL: Homography and epipolar-based consistent labeling for outdoor park surveillance. *Computer Vision and Image Understanding*, 111(1):21–42, 2008.

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

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of CVPR*, pages 2360–2367. IEEE, June 2010.

[7] M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Interactive person re-identification in TV series. In *Proc. of Int'l Workshop on Content Based Multimedia Indexing (CBMI)*, pages 1–6, June 2010.

[8] T. Gandhi and M. Trivedi. Panoramic Appearance Map (PAM) for Multi-camera Based Person Re-identification. In *Proc. of AVSS*, pages 78–78, Nov. 2006.

[9] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2007.

[10] D. Gray and H. Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Lecture Notes In Computer Science; Vol. 5302 - Proc. of ECCV: Part I*, page 262, 2008.

[11] L. Havasi, Z. Szlavik, and T. Sziranyi. Eigenwalks: walk detection and biometrics from symmetry patterns. In *Proc. of ICIP*, pages III–289, 2005.

[12] iLids. The Image library for intelligent detection systems, 2010. scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/i-lids/index.html.

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

[14] V. Leung, J. Orwell, and S. A. Velastin. Performance evaluation of re-acquisition methods for public transport surveillance. In *Proc. of ICCARV*, pages 705–712. IEEE, Dec. 2008.

[15] D. Maio, D. Maltoni, R. Cappelli, J. Wayman, and A. Jain. Fvc2000: Fingerprint verification competition. 24(3):402–412, March 2002.

[16] A.-T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *Proc. of AVSS*, 2007.

[17] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. of CVPR*, 2011.

[18] P. Over, G. Awad, J. Fiscus, B. Antonishek, A. F. Smeaton, W. Kraaij, and G. Quenot. Trecvid 2010 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. of TRECVID 2010*. NIST, USA, 2011.

[19] PETS. Dataset - Performance Evaluation of Tracking and Surveillance, 2009. http://www.cvg.rdg.ac.uk/PETS2009/.

[20] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *Proc. of CVPR*, pages 137–143, 1997.

[21] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. BMVC*, pages 21.1–11, 2010. doi:10.5244/C.24.21.

[22] N. Truongcong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. *Lecture Notes in Computer Science, Proceedings of ICIAP 2009*, N5716:p179–189, 2009.

[23] R. Vezzani and R. Cucchiara. Video surveillance online repository (visor): an integrated framework. *Multimedia Tools and Applications*, 50(2):359–380, Nov. 2010.

[24] R. Vezzani, C. Grana, and R. Cucchiara. Probabilistic people tracking with appearance models and occlusion classification: The ad-hoc system. *Pattern Recognition Letters*, 32(6):867–877, Apr. 2011.