Eurographics Italian Chapter Conference (2010)
E. Puppo, A. Brogni, and L. De Floriani (Editors)

3D Body Model Construction and Matching for Real Time People Re-Identification

D. Baltieri, R. Vezzani and R. Cucchiara

Dipartimento di Ingegneria dell'Informazione
University of Modena and Reggio Emilia
Via Vignolese, 905 - 41100 Modena - Italy

Abstract

Wide area video surveillance always requires to extract and integrate information coming from different cameras and views. Re-identification of people captured from different cameras or different views is one of most challenging problems. In this paper, we present a novel approach for people matching with vertices-based 3D human models. People are detected and tracked in each calibrated camera, and their silhouette, appearance, position and orientation are extracted and used to place, scale and orientate a 3D body model. Colour features are computed from the 2D appearance images and mapped to the 3D model vertices, generating the 3D model for each tracked person. A distance function between 3D models is defined in order to find matches among models belonging to the same person. This approach achieves robustness against partial occlusions, pose and viewpoint changes. A first experimental evaluation is conducted using images extracted from a real camera set-up.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking

1. Introduction

Surveillance of wide areas with several connected cameras integrated in the same automatic system is no more a chimera. The explosion of requests of intelligent surveillance systems for different scenarios such as people monitoring in public areas, smart homes, urban traffic control, mobile applications, and identity assessment for security and safety, leads research activity to explore many different dimensions in terms of both architectural issues and algorithms for scene recognition and interpretation. Thus, many synergic fields spanning from hardware embedded systems, sensor networks, computer architecture in one side and image processing, computer vision, pattern recognition and computer graphics in the other side are tightly integrated to cope with real-time surveillance applications [VC10].

Among the others, one of the most challenging task in these surveillance applications is to determine if a currently visible person has already been observed somewhere and sometime else in the network of cameras. In the computer vision community the problem is well known and called *People Re-identification*. This is a fundamental task for the analysis of long-term activities and behaviours by specific persons or to connect interrupted tracking. Algorithms have to be robust in challenging situations, like widely varying camera viewpoints and orientations, varying poses, rapid changes in clothes appearance, occlusions, and varying lighting conditions. The first studied re-identification problem was related to traffic analysis and vehicle tracking, where objects are rigid, move in well defined paths and have mostly uniform colours. Features as colour, size, speed, lane position are usually embedded in Bayesian frameworks [HR98]. People re-identification however requires more elaborate methods in order to cope with the widely varying degrees of freedom of a person's appearance.

Various algorithms have been proposed, based on the kind of data available: a first category of person re-identification methods rely on biometric techniques, such as face [BBF*10] or gait recognition [HSS05]. High resolution or PTZ cameras are required. Other approaches suppose easier operative conditions, calibrated cameras and precise knowledge of the geometry of the scene: the problem is then simplified by adding spatial and/or temporal constraints and reasoning in order to greatly reduce the candidate set [JSRS08] [MEB04] [VBC09]. Finally, most re-identification methods rely purely an appearance-based techniques [TCKA*10] [GBT07] [FBP*10] [HMSS08] [GT06] [AVBK10] [HJHG08] [LPB03]. Most of these method could be classified as single-shot algorithm, i.e. algorithm that associate pairs of images. [GBT07] evaluates some of the earliest and most simple methods: 1D Histograms of the entire appearance image, 3D Histograms, 3D correlograms, and hand local histograms and correlograms. These methods usually performs poorly in live situations but are often applied in easier operative conditions (as in [VBC09]). [TCKA*10] proposed instead a complex descriptor of a person silhouette called colour-position histogram: the silhouette is divided vertically in a fixed number of regularly spaced regions, and for each region a mean colour is computed. Then an algorithm based on spectral analysis and support vector machines provides a suitable classification for re-identification. Instead of regularly spaced regions [LPB03] uses the JSEG algorithm to segment a person images in different regions, and a colour and texture descriptor is computed for each region (a quantized HSV histograms and an edge energy descriptor); regions of different appearance are matched using the Integrated Region Matching algorithm. A more evolute approach was recently proposed by Farenzena *et al* [FBP*10] taking advantage of symmetry and asymmetry perceptual principles in order to subdivide the appearance images into 4 regions. For each region, three features describing complementary aspects of the human appearance are extracted: the overall chromatic content, spatial arrangement of colours into stable regions, and the presence of recurrent textures; a simple matching schema is then applied. Unlike the previous methods, [FBP*10] is a multiple-shot technique: descriptors from one or more images of the same person can be stored and used in the matching process. Interest point descriptors usually adopted for tracking or object recognition purposes such as SIFT, SURF and others have been effectively used for re-identification: [HMSS08] indicates an efficient variant of SURF and a fast technique for direct matching of interest points. [HJHG08] presents a similar method, based on SIFT descriptors and a previously trained Adaboost classifier. Lately, [AVBK10] proposed a more general (and multi-shot) framework for simultaneous tracking and re-detection by means of a grid cascade of dense region descriptors. In the paper, various descriptors have been evaluated, like SIFT, SURF and covariance matrices, and the latter are shown to outperform the formers. Finally, [GT06] proposed the concept of Panoramic Appearance Map to perform re-identification. This map is a compact signature of the appearance information of a person extracted from multiple cameras, and can be though of as the projection of a person appearance on the surface of a cylinder.

Our method is built on top of a previously developed tracking system [VC10] capable of extracting the silhouette of people and coarse data about the position, orientation and height of the person with respect to the viewing camera. Based on this data a hybrid 2D/3D re-identification algorithm is proposed. Using a simplified 3D model of the human body, simple dense descriptors are extracted from the appearance images and are projected on the surface of the 3D model. If multiple appearance images of the same track are available, the descriptors are merged together in the same model. Matching between people's appearances can be done in 2 ways: model-to-image, if only one image of the target is available, or model-to-model, if a complete 3D model of the target is available too. Our method has numerous advantage with respect to state of the art techniques for people re-identification: it uses regularly distributed dense descriptors, like [AVBK10], but organized in a 3-dimensional space, making our method truly robust to view changes and even partial occlusions. If more images of the appearance of the person are available, they can be efficiently stored in a fixed-size memory structure, keeping automatically only the more reliable informations.

2. 3D human model

Different graphic models have been adopted in the literature for 3D people tracking, motion capture, and posture analysis. These models are usually complex to deal with, requiring complex fitting techniques in order to obtain a perfect match between the 3D model posture and the real posture. Our goal is to create a 3D model, simple enough to be easily processed in real-time and, at the same time, informative enough to capture the complete appearance of a human figure. For this reason a monolithic 3D model is chosen.

In order to correctly construct the model, side, frontal and top different views of people were extracted from generic surveillance videos; then, for each view an average silhouette has been computed and used as reference for the manual creation of a 3D body model, exploiting state of the art computer graphic and 3d modeling software (in our case Lightwave 3D, v9.5), resulting in the model shown in fig. 1.

Then, the model surface has been sampled at different resolutions in order to generate sets of vertices which constitute our unspecific model. In this work we selected four different resolutions (i.e., different sampling density ρ), obtaining clouds with M_ρ=153, 628, 2026, and 10018 vertices respectively (see Fig.2).

The model described above is generic and does not contain any feature useful to differentiate and recognize people. Thus, for each person analysed by the system, a new instance of the generic model is created and integrated with a scale factor (to cope with different body builds) and appearance information, i.e., colour associated with each vertex. Calling Γ^p the p-th model instance, the specific model is then defined as:

$$\Gamma^p = \left\{h^p, \{V^p\}\right\} \tag{1}$$

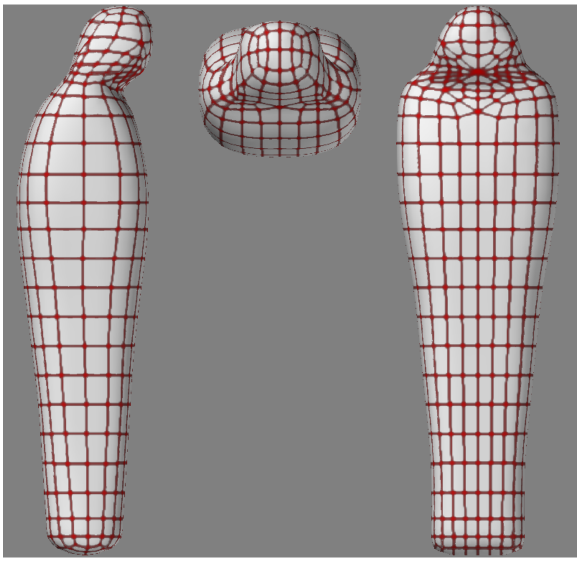

Figure 1: *3D model of a person*

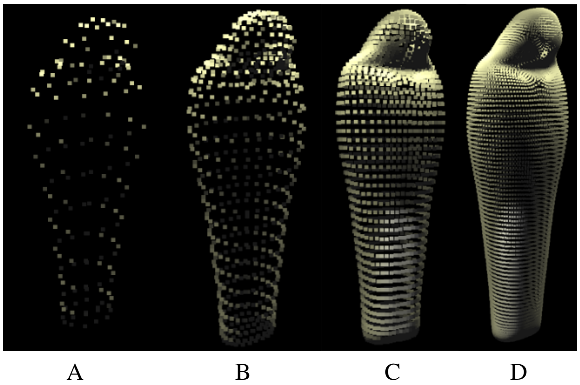

Figure 2: *Selected samplings of the 3D model with different density values (*ρ*=1, 4, 16, and 64)*

where h^p is the person height (extracted by the tracking module, and used as the scale factor for the 3d model) and $\{V^p\}$ is the vertex set. For each vertex $\{v_i^p \in V^p, i = 1 \dots M\}$ the following four features are computed and stored:

- $\vec{n}_i$: the normal vector computed at the vertex location from the original surface from which the vertices were sampled; this feature is static and pre-computed during the manual model creation;
- c_i: vertex colour in the RGB space for graphical interface purposes;
- $H(\cdot)_i$: a local HSV histogram which describes the colour appearance of the vertex neighbour; it is a normalized three dimensional histogram with 8 bins for the H channel and 4 bins for the S and V channels respectively;
- θ_i: the reliability value of the vertex data.

2.1. Model initialization

The tracking algorithm extracts the silhouette of the tracked person, its appearance image and its position, size and orientation with respect to the camera. In this preliminary work we are not interesting on the correct pose estimation of the monitored person. Thus, we assume that the person is walking unbowed and facing straight. The direction is estimated using the feet trajectory [CGPV05], and together with prior knowledge of the camera position and orientation it is possible to deduce with sufficient precision the position, orientation and size of the person for a correct alignment. Whenever the tracking system detects a new person in the scene, a corresponding specific model Γ^p is generated and tracking information together with camera calibration are exploited to initialize it.

Each vertex is projected onto the image plane of the appearance image (see Fig. 3).

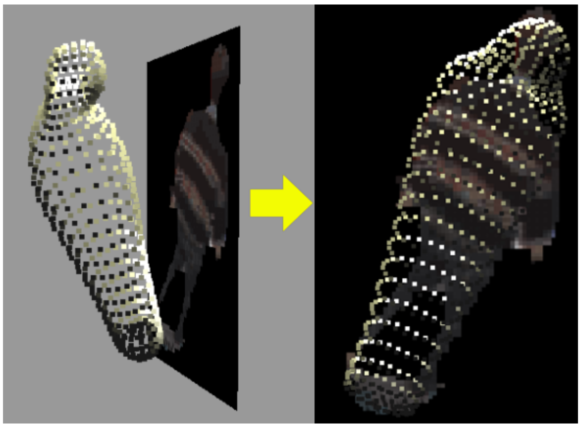

Figure 3: *Projection of the 3D model onto the appearance image*

The vertex colour is then initialized using the image pixel upon which the vertex was projected:

$$c_i = I\left(x(v_i), y(v_i)\right); \tag{2}$$

where I is the analysed frame, $x(v_i)$ and $y(v_i)$ are the image coordinates of the projection of the vertex v_i.

The reliability value is initialized with the result of the dot product between the vertex normal $\vec{n}_i$ and the image plane normal $\vec{p}$.

$$\theta_i = \vec{n}_i \cdot \vec{p} \tag{3}$$

The reason behind this is that data from front-viewed vertices and their surrounding surface is more reliable than that from lateral viewed vertices.

The histogram $H(\cdot)_i$ is computed on an image patch of size NxN centered around $(x(v_i), y(v_i))$.

The size of the patch, N, depends on the appearance image height (I_h), the vertex density (ρ) and a parameter α :

$$N = \frac{\alpha}{\sqrt{\rho}} I_h \qquad (4)$$

The vertices having no match with the current image (i.e. the vertices projected outside of the person silhouette) are iteratively initialized with a copy of the features of the nearest initialized vertex. The reliability value however, is set to $-1-\tau$.
An example of a resulting model can be seen in Fig. 4.

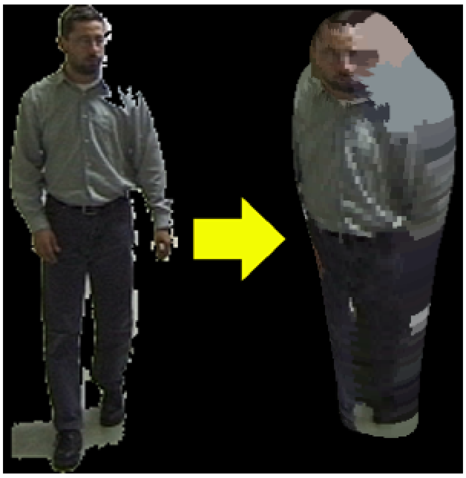

Figure 4: *Initialized 3D model of a person*

2.2. Model update

If the tracking algorithm can provide more than one image of the same person, the model can be further refined and updated in a way similar to the initialization of the model: for each appearance image available, the model vertices v_i^p are projected on it using the position and orientation parameters extracted by the tracking algorithm. Then, for each vertex successfully projected inside the silhouette of the person, a new feature vector is computed: the colour c_i^s of the pixel upon which the vertex was projected, a local HSV histogram $H(\cdot)_i^s$, computed on an image patch of size NxN (see Eq. 4) centered around the same pixel, and a new reliability value θ_i^s, computed as in the previous paragraph. The new feature vector is then averaged (or overwritten) with the existing one following the subsequent schema. Let θ_i^p be the reliability value of the feature vector stored in the model Γ^p:

- If $\theta_i^p < -1$ (i.e. the feature vector of the model vertex v_i^p was not initialized with real data) the new feature vector completely overwrite the one stored in the vertex v_i^p of the model Γ^p .
- If $\theta_i^p < 0$ and $\theta_i^s > 0$ (i.e. the feature vector of the model vertex v_i^p contains data initialized via symmetry, while the new feature vector is seen frontally) the new feature vector completely overwrite the one stored in the vertex v_i^p of the model Γ^p .
- If $\theta_i^p > 0$ and $\theta_i^s < 0$ (i.e the feature vector of the model vertex v_i^p contains data seen frontally, while the new feature vector contains symmetrically seen data) there is no update.
- If ($\theta_i^p > 0$ and $\theta_i^s > 0$) or ($\theta_i^p < 0$ and $\theta_i^s < 0$) (i.e. both the feature vector of the model vertex v_i^p and the new feature vector are frontally seen or symmetrically seen) a weighted average between the two feature vectors is computed (using θ_i^s and θ_i^p as weights) and stored in the vertex v_i^p of the model Γ^p .

The remaining vertices having $\theta_i^p < -1$ are reset and are iteratively initialized with a copy of the feature vector of the nearest initialized vertex. Since they have been propagated and are not related to a visible pixel, their reliability value is set to $-1-\tau$. Figure 5 shows an example of a model created from two images.

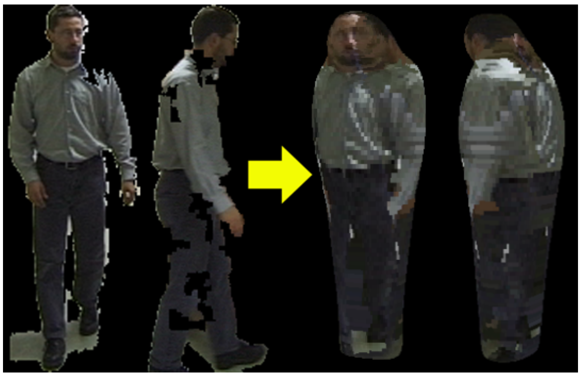

Figure 5: *Model Update: resulting model of a person obtained using two frames*

3. 3D models for re-identification

Given a library of specific person models, the re-identification problem is to detect new shows of the same person in the scene. Two different versions of the same method are here proposed for people re-identification, based on a model-to-image or a model-to-model matching respectively. The first one is to be used when only one image of the target is available, while the second one is best suited if the tracking algorithm has provided more than one appearance image of the target, so that a complete 3D model can be constructed. In the case of model-to-image matching the stored model is projected onto the appearance image of the target. For each vertex v_i^p of the model Γ^p that is projected inside the silhouette of the target person, the dot-product θ_i^s between the vertex normal $\vec{n}_i$ and the image plane normal $\vec{p}$ is computed, if the result is greater then a threshold value τ, a HSV localized histogram $H(\cdot)_i^s$ centered around the pixel in the projected vertex position $(x(v_i^p), y(v_i^p))$ is extracted (in the same way explained in the previous paragraphs) and its distance to the histogram contained in the feature vector of the model vertex v_i^p is finally computed. The formula used for the distance computation is the famous Bhattacharyya distance. The final matching score is the weighted average of all the computed distances.

$$d_H\left(H(\cdot)_i, H(\cdot)_j\right) = \sqrt{1 - \sum_{h,s,v} \sqrt{H(h,s,v)_i \cdot H(h,s,v)_j}} \tag{5}$$

$$D(\Gamma^p, S) = \frac{1}{K} \sum_{i=1...K} \left(d_H(H(\cdot)_i^p, H(\cdot)_i^s) \cdot f(\theta_i^p) \cdot f(\theta_i^s)\right) \tag{6}$$

$$f(\theta) = \begin{cases} \|\theta\|, & \text{if } \theta > -1 \\ \tau, & \text{if } \theta < -1 \end{cases} \tag{7}$$

Where K is the number of vertices projected inside the silhouette of the person.

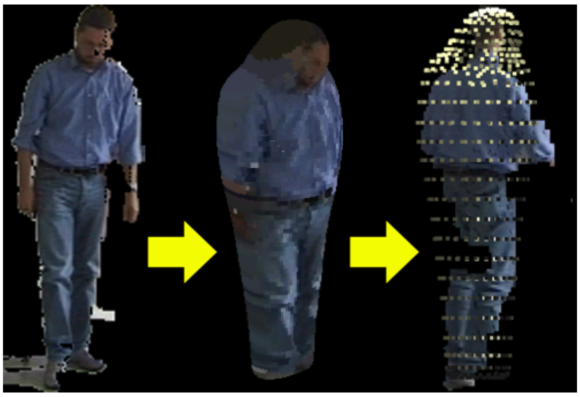

Figure 6: *Model-to-Image method*

In the model-to-model case, two models are available: the stored model Γ^p and a new model Γ^t computed from the images provided by the tracking algorithm. Calculating the matching score is more straightforward, being the geometrical models the same: for each pair of overlapping vertices, the Bhattacharyya distance between the HSV histograms (included in their respective feature vectors) is computed. Couples of vertices with at least one negative reliability value are discarded. The final score is computed as a weighted average of the remaining distances, using the product of the two reliability values as weight.

$$D(\Gamma^p, \Gamma^t) = \frac{1}{K} \sum_{i=1...K} \left(d_H(H(\cdot)_i^p, H(\cdot)_i^t) \cdot \theta_i^p \cdot \theta_i^t\right) \tag{8}$$

Where K is the number of vertices with both positive reliability values.

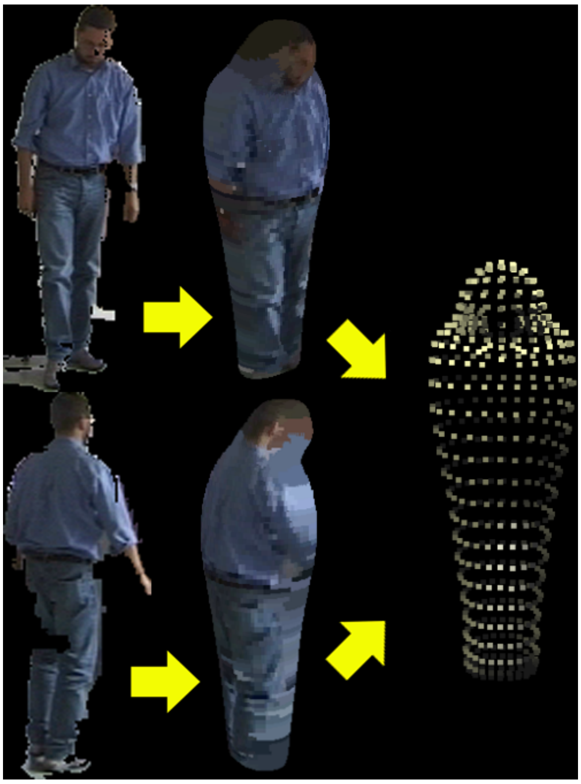

Figure 7: *Model-to-Model method*

A maximal criteria or a threshold can now be applied to select the best matching.

4. Experimental results

In order to test the proposed framework, we are creating a large dataset of videos containing walking people captured from different views. In this paper we used a subset of 28 videos (for a total of 20 people). The dataset contains, for each person, different snapshots from one or more cameras annotated with the person's silhouette, position, height and orientation. The calibration parameters of each camera are available as well.

Two replications of the experiments were run for each model. In each run two images for each person were randomly selected as training set and one for the test set. For each test image, a Nearest Neighbour criteria is adopted. Table 1 shows some preliminary results; the first column contains one of the two test images while the second and third

Test Image	1st Best Match [Distance]	2nd Best Match [Distance]
	0.38	
	0.25	0.54
	0.25	0.29
	0.21	0.59
	0.21	0.60
	0.08	0.60
	0.33	0.53
	0.49	**0.52**
	0.50	
	0.27	0.60
	0.25	0.48

Table 1: *Selected results from our tests using model B. The correct match is highlighted with bold font.*

column show the best and the second best match together with the corresponding distance values (candidates with a distance greater than 0.60 are omitted). In the reported experiments we adopted the model with 628 vertex and the other parameters were set as follows: α was set to $1/15$, τ to 0.5. In one case only the query returns an erroneous match.

To select the best density value for the model creation, we compared the four vertices density reported in section 2 (low, normal, high and very high): using the model with lowest density (most left in fig. 2) we obtained 78% of correct matches, while with higher densities (B and C in fig. 2) very good results have been achieved (in both cases resulting in about 92% correct matches). The highest density value, instead, (D in fig. 2) show worse results, with 65% correct matches only. This performance drop is probably due to the limited size of the input frames, which are not detailed enough to correctly extract all the model parameters. The density corresponding to a 628 vertex model is also optimal from the computational point of view; the model initialization step takes about 67 ms while the model matching 37ms on a single core of a Intel Core i5 750 CPU and a GeForce GT-240 video card. In case sufficiently precise data about the person position, size and orientation cannot be extracted by the tracking module, an additional alignment phase could be necessary, which will increase the time taken by the model matching phase. The prototype system has been developed in C++ based on the Imagelab framework [VC10] and the Ogre3D object-oriented graphics rendering engine [Ogr10] for graphical interface purposes and fast, hardware-accelerated geometrical transformations.

5. Conclusions and Future Works

We proposed a new and effective method for people re-identification. Differently from currently available solutions we exploit a 3D body model to spatially localize colour descriptors. In this way, occlusion and view dependencies are intrinsically solved. Some preliminary results have been obtained on a real dataset, demonstrating the efficacy of the proposal. We aim at integrating the proposed matching schema in our surveillance system [VC10], improving the model creation and updating with a completely automatic pose estimation and model fitting step. Moreover, other local descriptors such as the covariance matrix [TPM08] will be integrated in the vertex feature set in order to obtain a more expressive and powerful model.

The work is currently under development and improvement within the project THIS, 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.

References

[AVBK10] ALAHI A., VANDERGHEYNST P., BIERLAIRE M., KUNT M.: Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding 114*, 6 (2010), 624–640.

[BBF*10] BÄUML M., BERNARDIN K., FISCHER M., EKENEL H., STIEFELHAGEN R.: Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In *Proceedings of 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Boston* (2010).

[CGPV05] CUCCHIARA R., GRANA C., PRATI A., VEZZANI R.: Probabilistic posture classification for human behaviour analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 35*, 1 (Jan. 2005), 42–54.

[FBP*10] Farenzena M., Bazzani L., Perina A., Murino V., Cristani M.: Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (jun. 2010), pp. 2360–2367.

[GBT07] Gray D., Brennan S., Tao H.: Evaluating appearance models for recognition, reacquisition, and tracking. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)* (09/2007 2007).

[GT06] Gandhi T., Trivedi M.: Panoramic appearance map (pam) for multi-camera based person re-identification. In *IEEE International Conference on Video and Signal Based Surveillance, AVSS '06.* (nov. 2006), pp. 78 –78.

[HJHG08] Hu L., Jiang S., Huang Q., Gao W.: People re-detection using adaboost with sift and color correlogram. In *Proceedings of IEEE Int'l Conference on Image Processing* (2008), IEEE, pp. 1348–1351.

[HMSS08] Hamdoun O., Moutarde F., Stanciulescu B., Steux B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Second ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2008.* (sep. 2008), pp. 1–6.

[HR98] Huang T., Russell S.: Object identification: A bayesian analysis with application to traffic surveillance. *Artificial Intelligence 103* (1998), 1–17.

[HSS05] Havasi L., Szlavik Z., Sziranyi T.: Eigenwalks: walk detection and biometrics from symmetry patterns. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on* (sep. 2005), vol. 3, pp. III – 289–92.

[JSRS08] Javed O., Shafique K., Rasheed Z., Shah M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding 109*, 2 (2008), 146–162.

[LPB03] Lantagne M., Parizeau M., Bergevin R.: Vip: Vision tool for comparing images of people. *Vision Interface* (2003).

[MEB04] Makris D., Ellis T., Black J.: Bridging the gaps between cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2004.* (jun. 2004), vol. 2, pp. II–205 – II–210 Vol.2.

[Ogr10] Ogre3D.org: The ogre3d project homepage. http://www.ogre3d.org, Sept. 2010.

[TCKA*10] Truong Cong D. N., Khoudour L., Achard C., Meurie C., Lezoray O.: People re-identification by spectral classification of silhouettes. *Signal Process. 90*, 8 (2010), 2362–2374.

[TPM08] Tuzel O., Porikli F., Meer P.: Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 10 (oct. 2008), 1713–1727.

[VBC09] Vezzani R., Baltieri D., Cucchiara R.: Pathnodes integration of standalone particle filters for people tracking on distributed surveillance systems. In *Proceedings of the 15th International Conference on Image Analysis and Processing* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 404–413.

[VC10] Vezzani R., Cucchiara R.: Event driven software architecture for multi-camera and distributed surveillance research systems. In *Proceedings of the First IEEE Workshop on Camera Networks - CVPRW* (San Francisco, June 2010).