Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

Multi-Camera People Localization and Height Estimation using Multiple Birth-and-Death Dynamics

Ákos Utasi and Csaba Benedek

Hungarian Academy of Sciences, Computer and Automation Research Institute
Distributed Events Analysis Research Group
Kende u. 13-17. H-1111 Budapest, Hungary

Abstract. This paper presents a novel tool for localizing people in multi-camera environment using calibrated cameras. Additionally, we will estimate the height of each person in the scene. Currently, the presented method uses the human body silhouettes as input, but it can be easily modified to process other widely used object (*e.g.* head, leg, body) detection results. In the first step we project all the pixels of the silhouettes to the ground plane and to other parallel planes with different height. Then we extract our features, which are based on the physical properties of the 2-D image formation. The final configuration results (location and height) are obtained by an iterative stochastic optimization process, namely the multiple birth-and-death dynamics framework.

1 Introduction

Detecting and localizing people are key problems in many surveillance applications and are still challenging tasks in cluttered, crowded scenes due to the high occlusion rate caused by other people and static scene objects. Therefore, one object silhouette mask cannot be assumed to belong to only one person, and body masks can also break apart. Under such conditions single view localization or tracking might be impossible. The presented method is capable of accurately localizing individuals on the 3-D ground plane using multiple cameras. Hence, it can be used for many other high level machine vision tasks, such as scene understanding, multiple object tracking, or group/crowd behavior analysis. In addition, our method will also estimate the height of each individual. The proposed method assumes that the scene is monitored by multiple calibrated cameras, and the extracted human body silhouettes are available. These silhouettes are projected on the ground and multiple parallel planes. The presented method does not use any color or shape models for distinguishing multiple people in the scene. Instead, we will exploit the advantage of multiple cameras, and from the result of the multi-camera projection two similar geometric features are extracted in each 2-D position: one on the ground plane, and one on the other planes. Finally, the extracted features are used in a stochastic optimization process with geometric constraints to find the optimal configuration of multiple people.

The rest of the paper is organized as follows. In Sec. 2 we briefly present the related work in multi-camera people detection. The proposed method is discussed in Sec. 3. In Sec. 4 we evaluate our method using a public dataset. Finally, Sec. 5 concludes the paper.

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

2 Related work

In the last decades single-camera person detection and tracking has undergone a great evolution. See [1] for an extensive review of state-of-the-art methods. However, all of these methods have limited ability to handle crowded and cluttered scenes, where the occlusion rate is high. In such situations multi-view approaches provide a better solution, that can accurately estimate the position of multiple people. Mikic *et al.* [2] proposed a blob based approach (one object is represented by one blob on each view), where they estimated the 3-D centroid of an object by deriving a least squares solution of an over-determined linear system, where the measurements were the image coordinates of multiple views. [3] models the appearance (color) and locations of the people, to segment people on camera views. This helps the separation of foreground regions belonging to different objects. [4] extracts moving foreground blobs, and calculates the centroid of the blob's lowest pixels, which is projected on the ground plane. This information, in addition to the 2-D bounding box corners, is then used in a motion model. The method in [5] assumes that the objects are observed by multiple cameras at the head level. The ground plane is discretized into a grid, and from each grid position a rectangle (having the size of an average pedestrian) is projected to the camera views to model human occupancy. The method in [6] fuses evidence from multiple views to find image locations of scene points that are occupied by people. The homographic occupancy constraint is proposed, which fuses foreground likelihood information from multiple views to localize people on multiple parallel planes. This is performed by selecting one reference camera view and warping the likelihoods from the other views. Multi-plane projection is used to cope with special cases, when occupancy on the scene reference plane is intermittent (*e.g.* people running or jumping). In our method we also use multi-plane projection, but with a different purpose. We use the foreground masks from each camera, which are projected to the ground plane and to other parallel planes, and are used for feature extraction. Our hypothesis on the person's location and height is always a combination of evidences from two planes, the ground and the hypothetical head plane to form a discriminative feature. This is done by utilizing the 2-D image formation of the projected 3-D object. The method in [7] applies long–term statistical learning to make the spatial height distribution, which is used to estimate the height of a moving object. In our method such a long–term learning process is not needed, since the height of each person will be estimated during the optimization along with the position.

Another important issue is related to object modeling. *Direct* techniques construct the objects from primitives, like silhouette blobs [8] or segmented object parts. Although these methods can be fast, they may fail if the primitives cannot be reliably detected. On the other hand, *inverse methods* [9] assign a fitness value to each possible object configuration and an optimization process attempts to find the configuration with the highest confidence. In this way, flexible object appearance models can be adopted, and it is also straightforward to incorporate prior shape information and object interactions. However, search in the high dimensional population space has a high computational cost and the local maxima of the fitness function can mislead the optimization.

In the proposed model we attempt to merge the advantages of both low level and object level approaches. The applied Multiple Birth-and-Death (MBD) technique [9]

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

evolves the population of objects by alternating object proposition (*birth*) and removal (*death*) steps in a simulated annealing framework and the object verification follows the robust *inverse* modeling approach.

3 Proposed method

The input of the proposed method consists of human body silhouette masks extracted from multiple calibrated camera views (using Tsai's camera model[10]), monitoring the same scene. In our current implementation the foreground masks are obtained by first estimating a mixture of Gaussians (MoG) in each pixel [11], then the resulting models are used in the method of [12] without updating the model parameters. The main idea of our method is to project the extracted silhouettes both on the ground plane, and on the parallel plane shifted to the height of the person (see Fig. 1). This projection will create a distinct visual feature, and is visible from a virtual top viewpoint in the ground plane direction. However, no prior information of the persons height is known, and the height of different people in the scene may also be different. Therefore, we project the silhouette masks on multiple parallel planes with heights in the range of typical human height. In crowded scenes the overlapping rate is usually high, which would corrupt our hypothesis. We will solve this problem by fusing the projected results of multiple camera views on the same planes. The proposed method can be separated into the following three main steps and will be discussed in the subsequent sections in detail:

1. *Multi-plane projection:* The silhouettes are projected to the ground and to several parallel planes at different height.
2. *Feature extraction:* At each location of each plane we extract features that provide positive output for the real height and real location by using the physical properties of the 2-D image formation.
3. *Stochastic optimization:* We search for the optimal configuration in an iterative process using the extracted features and geometrical constraints.

3.1 Multi-plane projection

Let us denote by P_0 the ground plane, and by P_z the parallel plane above P_0 at distance z. In the first step of the proposed method we project the detected silhouettes to P_0 and

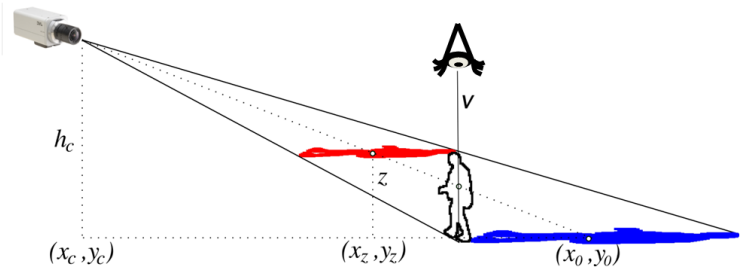
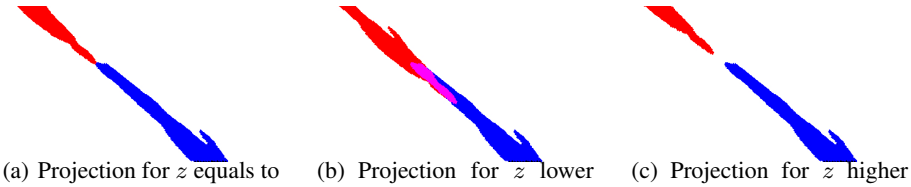


Fig. 1. Silhouettes are projected on the ground plane (blue) and on parallel planes (red).

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010



(a) Projection for z equals to the person's real height

(b) Projection for z lower than the person's real height

(c) Projection for z higher than the person's real height

Fig. 2. Our features are based on the 2-D image formation properties and on the multi-plane projection representation. The ground plane projection of one silhouette is marked with blue, and the P_z plane projection for three different z values (z is the distance from the ground) with red.

to different P_z planes (with different $z > 0$ offsets) by using the model of the calibrated cameras. As shown Fig. 1, this can be efficiently performed by projecting on P_0 only, then using the following relationship. Let (x_c, y_c) denote the position of an arbitrary camera and h_c its height, and let (x_0, y_0) denote the position of a selected point of the silhouette projected to the ground plane (*i.e.* $h_0 = 0$). Then the (x_z, y_z) position of the same point projected on a parallel plane at z height can be expressed as

$$x_z = x_0 - (x_0 - x_c)\, z/h_c \tag{1}$$

$$y_z = y_0 - (y_0 - y_c)\, z/h_c \tag{2}$$

In Fig. 1 and later in the text the projection of the silhouette to the P_0 ground plane is marked with blue, and to one P_z plane with red color.

3.2 Feature extraction

Our hypothesis on the location and height of a person is based on the physical properties of the 2-D image formation of a 3-D object. Consider the person with height h presented in Fig. 1, where we projected the silhouette on the P_0 ground plane (marked with blue) and the P_z plane with the height of the person (*i.e.* $z = h$, marked with red). Also consider the v vertical axis of the person that is perpendicular to the P_0 plane. We can observe that from this axis, the silhouette points projected to the $P_z|_{z=h}$ plane lie in the direction of the camera, while the silhouette print on P_0 is on the opposite side of v. For more precise investigations, in Fig. 2 the scene is visualized from a viewpoint above P_z, looking down on the ground plane in a perpendicular direction. Here, the silhouette prints from P_z and P_0 are projected to a common $x - y$ plane and jointly shown by red and blue colors, respectively (overlapping areas are purple). We can observe in Fig. 2(a), that if the height estimation is correct (*i.e.* $z = h$), the two prints just touch each other in the $p = (x, y)$ point which corresponds to the ground position of the person. However, if the z distance is underestimated (*i.e.* $z < h$), the two silhouette prints will overlap as shown in Fig. 2(b), and when the distance is overestimated (*i.e.* $z > h$), the silhouettes will move away, see Fig. 2(c).

Next, we derive a fitness function which evaluates the hypothesis of a proposed scene object with ground position $p = (x, y)$ and height h, using the information from

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

multiple cameras. Let (x_c^i, y_c^i) denote the projected position of the *i*th camera on the P_0 ground plane. We describe with angle $\varphi^i(p)$ the horizontal direction of the *i*th camera from p in the ground plane, calculated as:

$$\varphi^i(p) = \arctan\left(\frac{y - y_c^i}{x - x_c^i}\right) . \tag{3}$$

We will also use the definition of 'opposite' direction $\bar{\varphi}^i(p) = \varphi^i(p) + \pi$. The two directions are illustrated in Fig. 3(a).

Based on the above observations, an object hypothesis (x, y, h) is relevant according to the ith camera data if the following two conditions hold. ***Firstly***, we should find projected silhouette points on the P_0 plane (*i.e.* blue prints) around the $p = (x, y)$ point in the $\bar{\varphi}^i(p)$ direction, but penalize such silhouettes points in the $\varphi^i(p)$ direction of the same neighborhood. Considering these constraints, we define the $f_0^i(p)$ ground plane feature as:

$$f_0^i(p) = \frac{\mathbf{Area}\big(A_0^i \cap S(\bar{\varphi}^i(p), \Delta, p, r)\big) - \alpha \cdot \mathbf{Area}\big(A_0^i \cap S(\varphi^i(p), \Delta, p, r)\big)}{\mathbf{Area}\big(S(\bar{\varphi}^i(p), \Delta, p, r)\big)} , \tag{4}$$

where A_0^i denotes the set of silhouettes projected to plane P_0 using the *i*th camera model; $S(\bar{\varphi}, \Delta, p, r)$ and $S(\varphi, \Delta, p, r)$ denote the circular sectors with center p in the $[\bar{\varphi} - \Delta; \bar{\varphi} + \Delta]$ resp. $[\varphi - \Delta; \varphi + \Delta]$ angle range (marked with green on Fig. 3(a)), and r is a constant radius parameter being set a priori.

With notations similar to the previous case, we introduce the $f_z^i(p)$ feature on the P_z plane around the $p = (x, y)$ point in the $\varphi^i(p)$ direction as:

$$f_z^i(p) = \frac{\mathbf{Area}\big(A_z^i \cap S(\varphi^i(p), \Delta, p, r)\big) - \alpha \cdot \mathbf{Area}\big(A_z^i \cap S(\bar{\varphi}^i(p), \Delta, p, r)\big\}}{\mathbf{Area}\big(S(\varphi^i(p), \Delta, p, r)\big)} . \tag{5}$$

Both $f_0^i(p)$ and $f_z^i(p)$ are then truncated to take values in the $[0, \bar{f}]$ range, and are normalized by $\bar{f}$. Here, $\bar{f}$ controls the area ratio required to produce the maximal output.

If the object defined by the (x, y, h) parameter set is fully visible for the ith camera, both the $f_0^i(p)$ and $f_z^i(p)$ features should have *high* values in point $p = (x, y)$ and height $z = h$. Unfortunately in the available views, some of the legs or heads may be partially or completely occluded by other pedestrians or static scene objects, which will strongly corrupt the feature values. Although $f_0^i(p)$ and $f_z^i(p)$ features are weak in the individual cameras, we can construct a strong classifier if we consider all the camera data simultaneously and calculate the product of the average of the calculated feature values over the different views, *i.e.*

$$f(p, z) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} f_0^i(p) \times \frac{1}{N}\sum_{i=1}^{N} f_z^i(p)} . \tag{6}$$

After the above feature definition, finding all the pedestrians in the scene is done by a global optimization process. Since the number of people is also unknown, and each person should be characterized by its x, y and h parameters, the configuration space has a high dimension, therefore an efficient optimization technique should be applied.

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

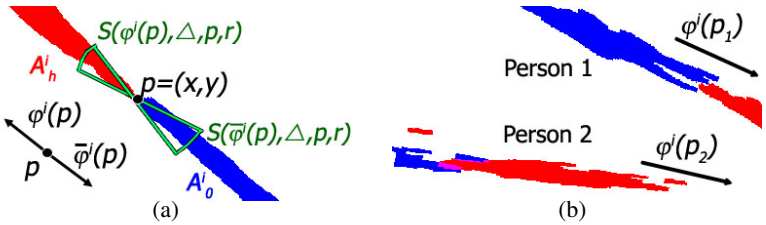


Fig. 3. (a) Notations and areas used for the calculation of the $f_0^i(p)$ and $f_z^i(p)$ features. (b) Silhouette prints to P_0 and P_z at a given z distance from a scenario with two people. Person 1's height has been accurately found ($h_1 = z$), however Person 2's one is underestimated ($z < h_2$).

3.3 Marked Point Process model

Our goal is to detect and separate the people in the scene, and provide their position and height parameters. For this reason, we will use a simplified object model: we describe the people by their bounding cylinders in the 3-D space. Let us assume that the ground plane is flat and the people are standing on the ground. Thus, a given object-cylinder u is defined by its $x(u)$ and $y(u)$ coordinates in the ground plane and the $h(u)$ height of the cylinder, as shown in Fig. 4(a).

Let $\mathcal{H}$ be the space of u objects. The Ω configuration space is defined as [9]:

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \quad \Omega_n = \left\{ \{u_1, \ldots, u_n\} \in \mathcal{H}^n \right\}. \tag{7}$$

Let ω denote an arbitrary object configuration $\{u_1, \ldots, u_n\}$ in Ω. We define a $\sim$ neighborhood relation in $\mathcal{H}$: $u \sim v$ if the cylinders intersect. We refer to the global input data with $\mathcal{D}$ in the model which consists in the foreground silhouettes in all camera views and the camera matrices.

We introduce a non-homogeneous input-dependent energy function on the configuration space: $\Phi_{\mathcal{D}}(\omega)$, which assigns a *negative likelihood* value to each possible object

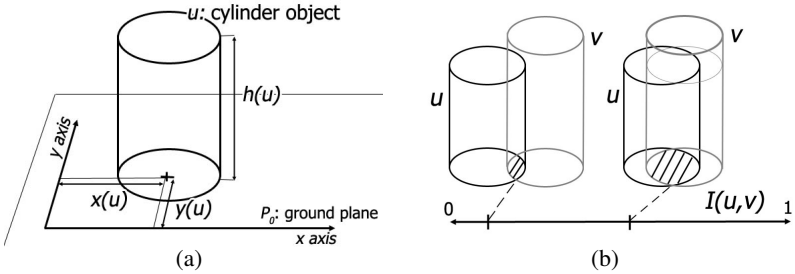


Fig. 4. (a) Cylinder objects are used to model persons in the 3-D space. Their ground plane position and height will be estimated. (b) Intersection of cylinders in the 3-D space is used as geometrical constraint in the object model.

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

population. The energy is divided into data dependent ($J_{\mathcal{D}}$) and prior (I) parts:

$$\Phi_{\mathcal{D}}(\omega) = \sum_{u \in \omega} J_{\mathcal{D}}(u) + \gamma \cdot \sum_{\substack{u,v \in \omega \\ u \sim v}} I(u,v)\,, \tag{8}$$

where $J_{\mathcal{D}}(u) \in [-1,1]$, $I(u,v) \in [0,1]$ and γ is a weighting factor between the two terms. We derive the optimal object configuration as the maximum likelihood configuration estimate, which can be obtained as $\omega_{\mathrm{ML}} = \arg\min_{\omega \in \Omega} \left[\Phi_{\mathcal{D}}(\omega)\right]$.

The next key task is to define the I prior and $J_{\mathcal{D}}$ data-based potential functions appropriately so that the ω_{ML} configuration efficiently estimates the true group of people in the scene. First of all, we have to avoid configurations which contain many objects in the same or strongly overlapping positions. Therefore, the $I(u,v)$ *interaction* potentials realize a prior geometrical constraint: they penalize intersection between different object cylinders in the 3-D model space (see Fig. 4(b)) :

$$I(u,v) = \mathbf{Area}\big(u \cap v\big) / \mathbf{Area}\big(u \cup v\big)\,. \tag{9}$$

On the other hand, the $J_{\mathcal{D}}(u)$ *unary* potential characterizes a proposed object candidate segment $u = (x, y, h)$ depending on the local image data, but independent of other objects of the population. Cylinders with negative unary potentials are called *attractive objects*. Considering (8) we can observe that the optimal population should consist of attractive objects exclusively: if $J_{\mathcal{D}}(u) > 0$, removing u from the configuration results in a lower $\Phi_{\mathcal{D}}(\omega)$ global energy.

At this point we utilize the $f_u = f(p(u), h(u))|_{p(u)=(x(u),y(u))}$ feature in the MPP model, which was introduced in Sec. 3.2. Let us remember, that the f_u fitness function evaluates a person-hypothesis for u in the multi-view scene, so that 'high' f_u values correspond to efficient object candidates. For this reason, we project the feature domain to $[-1,1]$ with a monotonously decreasing function (see also Fig. 5):

$$J_{\mathcal{D}}(u) = Q(f_u, d_0, D) = \begin{cases} \left(1 - \frac{f_u}{d_0}\right) & \text{if } \; f_u < d_0 \\ \exp\left(-\frac{f_u - d_0}{D}\right) - 1 & \text{if } \; f_u \geq d_0 \end{cases} \tag{10}$$

where d_0 and D are parameters. Consequently, object u is attractive according to the $J_{\mathcal{D}}(u)$ term iff $f_u > d_0$, while D performs data-normalization.

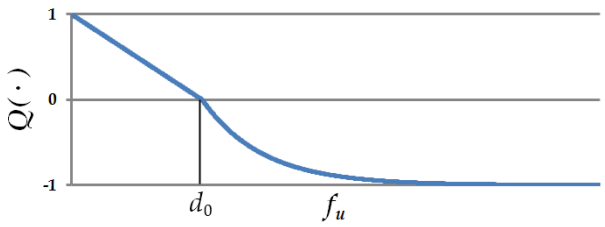


Fig. 5. Plot of the $Q(f_u, d_0, D)$ function

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

3.4 Optimization by multiple birth-and-death dynamics

We estimate the optimal object configuration by the Multiple Birth and Death Algorithm [9] as follows:

Initialization: start with an empty population $\omega = \emptyset$, and fit a 2-D pixel lattice to the P_0 ground plane. Let s denote a single pixel of this lattice.

Main program: set the birth rate b_0, initialize the inverse temperature parameter $\beta = \beta_0$ and the discretization step $\delta = \delta_0$, and alternate birth and death steps.

1. *Birth step*: Visit all pixels on the ground plane lattice one after another. At each pixel s,if there is no object with ground center s in the current configuration ω, choose birth with probability δb_0.
   If birth is chosen at s: generate a new object u with ground center $[x(u), y(u)] := s$, and set the height parameter $h(u)$ randomly between prescribed maximal and minimal height values. Finally, add u to the current configuration ω.
2. *Death step*: Consider the configuration of objects $\omega = \{u_1, \ldots, u_n\}$ and sort it by decreasing values of $J_{\mathcal{D}}(u)$. For each object u taken in this order, compute $\Delta\Phi_\omega(u) = \Phi_{\mathcal{D}}(\omega/\{u\}) - \Phi_{\mathcal{D}}(\omega)$, derive the *death rate* as follows:

$$d_\omega(u) = \frac{\delta a_\omega(u)}{1 + \delta a_\omega(u)}, \text{ with } a_\omega(u) = e^{-\beta \cdot \Delta\Phi_\omega(u)}$$

   and remove u from ω with probability $d_\omega(u)$.

Convergence test: if the process has not converged yet, increase the inverse temperature β and decrease the discretization step δ with a geometric scheme, and go back to the birth step. The convergence is obtained when all the objects added during the birth step, and only these ones, have been killed during the death step.

4 Experiments

To test our method we used the *City center* images of the PETS 2009 dataset [13] containing 400 video frames, and selected cameras with large fields of view (View_001, View_002, and View_003). In our experiments the projections were limited to a manually selected rectangular area on the ground plane, visible from all cameras. The MoG background model was defined in the CIE L*U*V* color space, and after the parameter estimation process the covariances were manually increased to have a minimum value of 25.0 (chroma channels) or 49.0 (luma channel) to reduce the effects of cast shadow. Finally, to separate the foreground from the background the technique of [12] was used with the following settings: modality parameter $T = 0.6$, matching criterion $I = 3.0$.

In the feature extraction step (Sec. 3.2) we assumed that $r = 25cm$, Δ was set to constant $30°$, the penalty parameter to $\alpha = 1.0$, and the area ratio threshold to $\bar{f} = 0.75$. To set the parameters of the optimization process we assumed that at least one view should correctly contain the feet and another one the head of a person, which implies a $d_0 = 1/3$ threshold for object candidate acceptance. However, due to the noisy foreground masks, in our experiments we used a less restrictive value of $d_0 = 0.28$. D was set to constant 8, and we assumed a minimum distance constraint of $50cm$ between two

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

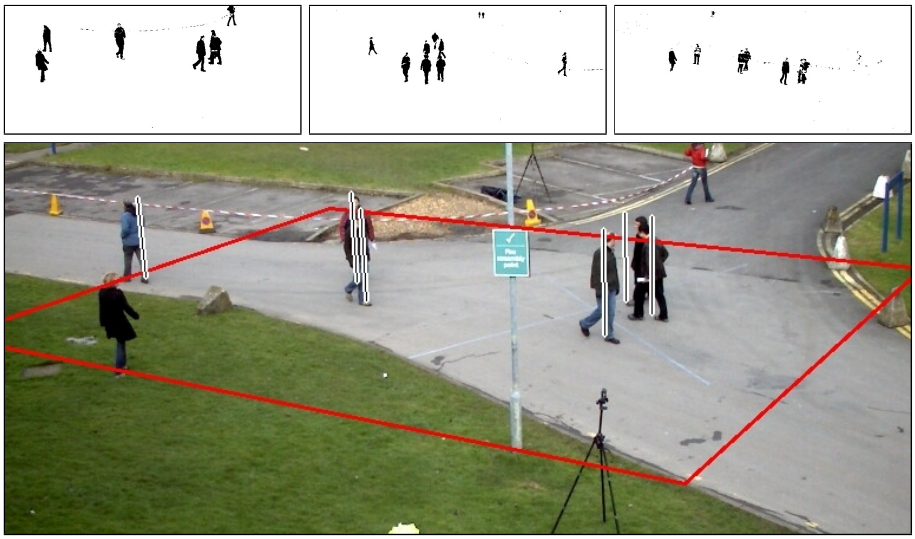


Fig. 6. Top: result of the foreground-background separation. Bottom: estimated ground position and height of each person represented by a line. The monitored area is represented by a red rectangle.

people (*i.e.* the radius of the cylinder in Fig. 4(a)). As for the parameters of the Multiple Birth and Death optimization process, we followed the guidelines provided in [9], and used $\delta_0 = 20000$, $\beta_0 = 50$, and geometric cooling factors $1/0.96$. For each video frame we limited the optimization process to a maximum of 20 iterations, and did not use the result for the subsequent time step. For visualizing the results, we backprojected the estimated positions on the first camera view and draw a line between the ground plane and the estimated height (see Fig. 6 bottom), the monitored area boundary is represented by a red rectangle. Figure 6 top contains the results of the foreground-background separation. Finally, we visually evaluated the inaccuracy rate of the results (*i.e.* of positive detections with under- or overestimated height, being 6.27%), and we also calculated the rate of missed detections (being 1.75%). Further experimental results may be found at `http://web.eee.sztaki.hu/~ucu/vs10-location-results.avi`.

5 Conclusion

In this paper we presented a novel method to localize people in multiple calibrated cameras. For this tasks we extracted a feature, which is based on the physical properties of the 2-D image formation, and produces high response (evidence) for the real position and height of a person. To get a robust tool for cluttered scenes with high occlusion rate, our approach fuses evidences from multi-plane projections from each camera. Finally, the positions and heights are estimated by a constrained optimization process, namely the Multiple Birth-and-Death Dynamics. In the current implementation we use foreground-background separation [12] to extract foreground pixels. For evaluation we

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/

Author manuscript, published in Workshop on Visual Surveillance, pp. 74-83, Queenstown, New Zeland, 2010

used the images of a public outdoor dataset, containing three camera views. According to our tests, the proposed method produces accurate estimation, even in cluttered environment, where full or partial occlusion is present. In the future we will investigate the effects of the different parameter settings of the feature extraction and the optimization steps. Moreover, we will examine the advance of using the optimization result for the estimation process of the subsequent time step. Another possible improvement might be the use of a robust body part detector (*e.g.* [14]) for creating evidence. This can be easily integrated in the proposed algorithm with minimal modification.

Acknowledgments

This work was partially supported by the THIS (Transport Hub Intelligent video System) project of the EU.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys **38** (2006) 13
2. Mikic, I., Santini, S., Jain, R.: Video processing and integration from multiple cameras. In: Proc. of the Image Understanding Workshop. (1998) 183–187
3. Mittal, A., Davis, L.S.: M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. Int. J. of Computer Vision **51** (2002) 189–203
4. Kang, J., Cohen, I., Medioni, G.: Tracking people in crowded scenes across multiple cameras. In: Proc. of the Asian Conf. on Computer Vision. (2004)
5. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. on Pattern Analysis and Machine Intelligence **30** (2008) 267–282
6. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. IEEE Trans. on Pattern Analysis and Machine Intelligence **31** (2009) 505–519
7. Havasi, L., Szlávik, Z.: Using location and motion statistics for the localization of moving objects in multiple camera surveillance videos. In: Proc. of the IEEE Int. Workshop on Visual Surveillance. (2009)
8. Benedek, Cs., Szirányi, T.: Bayesian foreground and shadow detection in uncertain frame rate surveillance videos. IEEE Trans. on Image Processing **17** (2008) 608–621
9. Descombes, X., Minlos, R., Zhizhina, E.: Object extraction using a stochastic birth-and-death dynamics in continuum. J. of Math. Imaging and Vision **33** (2009) 347–359
10. Tsai, R.Y.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. IEEE J. of Robotics and Automation **3** (1987) 323–344
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. of Royal Statistical Society, Series B **39** (1977) 1–38
12. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 747–757
13. PETS: Dataset - Performance Evaluation of Tracking and Surveillance (2009) `http://www.cvg.rdg.ac.uk/PETS2009/a.html`.
14. Wu, B., Nevatia, R.: Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. Int. J. of Computer Vision **82** (2009) 185–204

Document version of the MTA SZTAKI Publication Repository, http://eprints.sztaki.hu/