A 3-D Marked Point Process Model for Multi-View People Detection

Ákos Utasi and Csaba Benedek
Computer and Automation Research Institute, Hungarian Academy of Sciences
Kende u. 13-17. H-1111 Budapest, Hungary
{utasi,bcsaba}@sztaki.hu

Abstract

In this paper we introduce a probabilistic approach on multiple person localization using multiple calibrated camera views. People present in the scene are approximated by a population of cylinder objects in the 3-D world coordinate system, which is a realization of a Marked Point Process. The observation model is based on the projection of the pixels of the obtained motion masks in the different camera images to the ground plane and to other parallel planes with different height. The proposed pixel-level feature is based on physical properties of the 2-D image formation process and can accurately localize the leg position on the ground plane and estimate the height of the people, even if the area of interest is only a part of the scene, meanwhile silhouettes from irrelevant outside motions may significantly overlap with the monitored region in some of the camera views.

We introduce an energy function, which contains a data term calculated from the extracted features and a geometrical constraint term modeling the distance between two people. The final configuration results (location and height) are obtained by an iterative stochastic energy optimization process, called the Multiple Birth and Death dynamics. The proposed approached is compared to a recent state-of-the-art technique in a publicly available dataset and its advantages are quantitatively demonstrated.

1. Introduction

Detecting and localizing people are key issues in many surveillance applications, such as person tracking or people counting. The task is still challenging in cluttered, crowded scenes due to the high occlusion rate between the different moving and static scene objects. Background subtraction is one of the basic tools to deal with the problem, however, in a crowded scenario a given object silhouette blob in the foreground mask may belong to more than one person, and due to noise and occlusion, body masks can also break apart. Under such conditions single view localization approaches might be inefficient: a straightforward improvement is to utilize images of different cameras from different viewpoints in parallel. 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 people counting. 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 foreground masks are available. The foreground pixels 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 pixel-level features are extracted in each 2-D position: one on the ground plane, and one on each head plane. Both features collect evidence for existence of people on the ground. 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.

2. Related Work

People localization is the first key step in many machine vision applications, such as person tracking or counting. In the last decades single-camera person detection and tracking has undergone a great evolution. See [17] 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.* [11] 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. [12] 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. [9] 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. All the above methods attempt to extract complete object shapes, which is inefficient in cluttered environment where the objects break apart or when the occlusion rate is high. Therefore, the features we use in the proposed method are defined at pixel-level. The method in [4] 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 [10] 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 pixel-level 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 [8] applies long–term statistical learning to make the spatial height distribution, which is used to estimate the an object's height. In our method such a long–term process is not needed, since the person's height is estimated during the optimization along with the position.

We also need to deal with object modeling. *Direct* techniques construct the objects from primitives, like silhouette blobs [1] 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*, such as Marked Point Processes (MPP) [3] 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 [6] a single view MPP model is developed to detect and count people in crowded scenes. The model couples a spatial stochastic process governing number and placement of individuals with a conditional mark process for selecting body shape. However, limitations from the monocular approach results in difficulties in strongly crowded scenarios, where the overlapping rate is high. On the contrary, we optimize the objects in the 3-D real world space instead of the 2-D shapes in the individual camera views similarly to [7]. The main difference from the latter approach lies in the data model construction as our proposed pixel-level feature focuses on the accurate extraction of the foot and head point of the people, instead of considering the whole silhouettes which may be corrupted by overlapping or disruption effects. This property results in efficient localization, even if the area of interest is only a part of the scene, while silhouettes from irrelevant outside motions significantly overlap with the monitored region in some of the camera views. On the other hand, instead of utilizing the conventional Reverse Jump Markov Chain Monte Carlo (RJMCMC) optimization method, which tends to be sensitive to false local maxima resulting in ghost effects [7], we apply the recently proposed Multiple Birth-and-Death (MBD) technique [3] which is by nature less influenced by the above artifact. The population of objects is evolved by alternating multiple object proposition (*birth*) and removal (*death*) steps in a simulated annealing framework and the object verification follows the robust *inverse* modeling approach. In contrast to RJMCMC, in MBD each birth step consists in adding several random objects to the current configuration. In addition, there is no rejection during the birth move, therefore high energetic objects can be still added independently of the temperature parameter - this property prevents the algorithm from being stucked in ghost objects.

3. Proposed Method

The input of the proposed method consists of foreground masks extracted from multiple calibrated camera views (using Tsai's camera model [15]), monitoring the same scene. In our current implementation the masks are obtained by using a mixture of Gaussians (MoG) background model. The main idea of our method is to project the extracted foreground pixels 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 birds-eye 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 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 pixel-level 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 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

$$x_z = x_0 - (x_0 - x_c)\, z/h_c \quad (1)$$

$$y_z = y_0 - (y_0 - y_c)\, z/h_c \quad (2)$$

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. Pixel-Level 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).

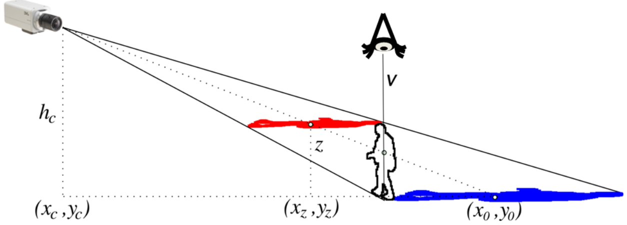


Figure 1. Silhouettes are projected on the ground plane (blue) and on parallel planes (red).

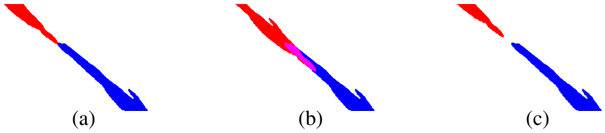

(a) (b) (c)

Figure 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. (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.

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

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

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.

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,

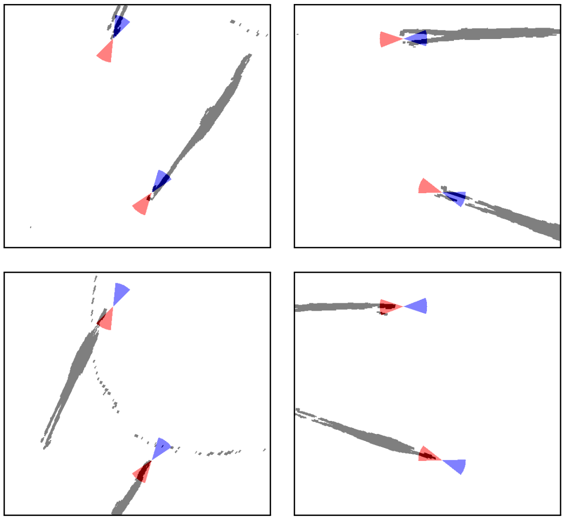


Figure 3. Top: ground feature extraction from two camera views. Bottom: head feature extraction from two camera views on plane $P_{z=168\text{cm}}$. The sector $S^i(p)$ in the camera's direction $\varphi^i(p)$ is denoted by red, in the opposite direction $\bar{\varphi}^i(p)$ by blue color.

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 in the following way. Let us denote by A_0^i the set of foreground pixels projected to plane P_0 using the ith camera model; $\bar{S}^i(p) = S(\bar{\varphi}^i(p), \varDelta, p, r)$ and $S^i(p) = S(\varphi^i(p), \varDelta, p, r)$ denote the circular sectors with center p in the $[\bar{\varphi}^i(p) - \varDelta; \bar{\varphi}^i(p) + \varDelta]$ resp. $[\varphi^i(p) - \varDelta; \varphi^i(p) + \varDelta]$ angle range (marked with blue resp. red on Fig. 3), and r is a constant radius parameter being set a priori. Then the $f_0^i(p)$ feature is calculated as:

$$f_0^i(p) = \frac{\mathbf{A}\big(A_0^i \cap \bar{S}^i(p)\big) - \alpha \cdot \mathbf{A}\big(A_0^i \cap S^i(p)\big)}{\mathbf{A}\big(\bar{S}^i(p)\big)}\,, \qquad (4)$$

where $\mathbf{A}$ denotes the area. 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:

$$f_z^i(p) = \frac{\mathbf{A}\big(A_z^i \cap S^i(p)\big) - \alpha \cdot \mathbf{A}\big(A_z^i \cap \bar{S}^i(p)\big)}{\mathbf{A}\big(S^i(p)\big)}\,. \qquad (5)$$

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 plane 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 feature if we fuse all the camera data by calculating the product of the average of the calculated feature values over the different views, *i.e.*

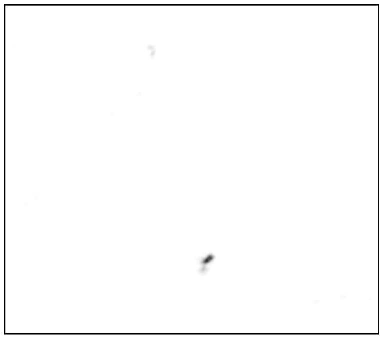


Figure 4. $f(\cdot, z = 168\text{cm})$ strong features calculated by fusing the features of Fig. 3 by (6). Low intensity pixels indicate the most probable person positions.

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

Fig. 3.2 demonstrates the output of (6) on plane $P_{z=168\text{cm}}$ using the features presented in Fig. 3.

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.

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 use a simplified object model: we describe the people by cylinders with fixed R radius in the 3-D world coordinate system. Let us assume that the ground is flat and the people are standing on it. The area of interest which we monitor in the P_0 ground plane is rectangular. 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. 5(a).

In the implementation, we use a discrete space of the objects: we discretize the area of interest in P_0 into $S_W \times S_H$ locations corresponding to a regular grid, and also round the person heights to integers measured in cm. Therefore, the object space $\mathcal{H}$ can be obtained as $\mathcal{H} = [1, \dots, S_W] \times [1, \dots, S_H] \times [h_{\min}, \dots, h_{\max}]$.

We aim to extract a configuration of a finite number of cylinder objects in the scene. Thus the Ω configuration space is defined as:

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \quad \Omega_n = \big\{\{u_1, \dots, u_n\} \in \mathcal{H}^n\big\}\,. \qquad (7)$$

Let ω denote an arbitrary object configuration $\{u_1, \ldots, u_n\}$. 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 population. The energy is divided into data dependent ($J_\mathcal{D}$) and prior (I) parts:

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

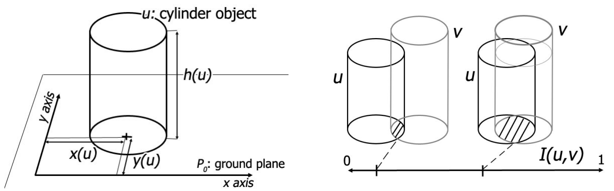
where $J_\mathcal{D}(u) \in [-1, 1]$, $I\,(u, v) \in [0, 1]$ and γ is a weighting factor between the two terms. We derive the optimal object configuration as the maximum likelihood configuration estimate, which can be obtained as

$$\omega_{\text{ML}} = \operatorname*{argmin}_{\omega \in \Omega} \left[\Phi_\mathcal{D}(\omega)\right] \qquad (9)$$

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. 5(b)) :

$$I(u,v) = \mathbf{Area}\big(u \cap v\big) / \mathbf{Area}\big(u \cup v\big)\ . \qquad (10)$$

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.



(a) (b)

Figure 5. (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.

At this point we utilize the $f_u = f(p(u), h(u))$ feature at ground point $p(u) = (x(u), y(u))$ 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. 6):

$$\begin{aligned} J_\mathcal{D}(u) =& Q(f_u, d_0, D) = \\ =& \begin{cases} \left(1 - \frac{f_u}{d_0}\right) & \text{if } \ f_u < d_0 \\ \exp\left(-\frac{f_u - d_0}{D}\right) - 1 & \text{if } \ f_u \geq d_0 \end{cases} \end{aligned} \qquad (11)$$

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. Thus d_0 parameter defines the minimal feature value required for object acceptance.

3.4. Optimization

Even with the discretization of the $x(u)$, $y(u)$ and $h(u)$ object descriptors, and prescribing as constraint at most one person in a given ground position, the cardinality of the population space is exponential function of the number of possible locations. For example, we used for the PETS dataset 609×745 locations in P_0 and 55 different height values (between 155 and 210 cm, with to 1 cm accuracy), which yields $(55 + 1)^{609 \times 745}$ different configurations - as each location may be empty, or contain a person with arbitrary height. Thus exhaustive search cannot be fulfilled for minimizing (9), instead of this, we should adopt techniques which can efficiently sample the configuration space.

In previous MPP applications, various optimization methods have been utilized [5], mainly implementing an iterative process which consists of object proposition (birth) and removal (death) steps. The most widely used approach has been the RJMCMC technique [7], where to the birth step, moves are added such as split, translate, rotate, etc. The main limitation is that here each iteration consists in perturbing one or a couple of objects and the rejection rate induces a huge computation time. A quicker algorithm - called Multiple Birth and Death (MBD) - has been proposed in [3], which enables multiple perturbations in parallel, resulting in increased speed of convergence and simplicity of implementation. Note that a graph cut based method (Multiple Birth and Cut, MBC) has been published very recently

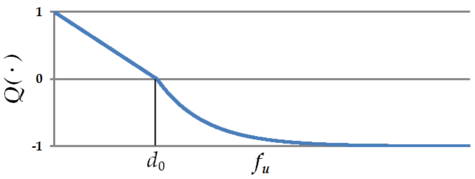


Figure 6. Plot of the $Q(f_u, d_0, D)$ function

[5], reporting slight quality improvements, but higher computational cost than MBD. For choosing a trade-of between speed and quality we have adapted the MBD optimization to our MPP model. The steps are as follows:

Initialization: start with an empty population $\omega = \emptyset$, and fit a 2-D pixel lattice to the P_0 ground plane - one pixel s for each location of interest.

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

We have compared our approach to the Probabilistic Occupancy Map (POM) technique [4], which is a state-of-the-art method with similar purposes[1]. This procedure estimates the marginal probabilities of presence of individuals at every location in an area of interest under a simple appearance model, given binary images corresponding to the result of a background-subtraction from different viewpoints. The appearance model is parametrized by a family of rectangles which approximate the silhouettes of individuals standing at every location of interest, from every point of view.

For the evaluation of the two methods 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) and we used an area of interest of size 12.2 m × 14.9 m, which is visible from all three cameras.

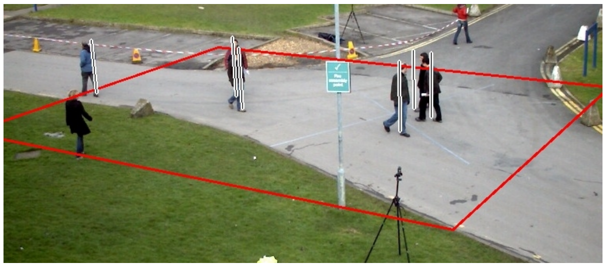


Figure 7. Estimated ground position and height of each person represented by a line. The monitored area is represented by a red rectangle.

For foreground extraction we used a MoG background model in the CIE L*u*v* color space. First, the MoG parameters were estimated by offline training [2], then the covariances were manually increased to have a minimum value of 25.0 (chroma channels) and 49.0 (luma channel) to reduce the effects of cast shadow. Finally, to separate the foreground from the background the technique of [14] was used with the following settings: modality parameter $T = 0.6$, matching criterion $I = 3.0$, the background model was not updated. During the evaluation of the POM, we manually masked out the regions on each camera, which do not belong the volume of interest defined by a rectangular cuboid[2]. Our method does not require such region masking, therefore this step was neglected in its evaluation.

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. 7), the monitored area boundary is represented by a red rectangle. Finally, we visually evaluated the results by carefully counting:

1. Missed Detection (**MD**, see Fig. 8(a)): $\#\{$human bodies, that were not detected$\}$;

2. False Detection (**FD**, see Fig. 8(b)): $\#\{$false detections appear at positions, which are not occupied by a person$\}$;

3. Multiple Instances (**MI**, see Fig. 8(c)): $\#\{$people localized multiple times in the same video frame at different positions$\}$.

POM: To evaluate the POM method we used the discretization parameter proposed in [4], thus the area of interest was discretized it into $G = 2940$ locations, corresponding to a regular grid with a 25 cm resolution. Since POM

[1] Executable application of the technique is freely available at http://cvlab.epfl.ch/software/pom/

[2] This step was performed to improve the stability of the algorithm, and was advised by the authors.

uses fixed human height estimates, from the camera calibration, we defined for each camera a family of rectangular shapes which correspond to crude human silhouettes of height 175 cm and width 50 cm located at every position on the grid. The generative method outputs grid position occupancy probabilities, therefore in our evaluation we simply use a threshold $\mathbf{T}$ to classify people locations.

We used the publicly available source code for the optimization. Most of the noticed artifacts of POM have been resulted by the fact that people are sometimes observed outside that area of interest, however the projections of their foreground silhouette masks in the different camera views overlap with monitored image regions. This limitation - which is more robustly handled in our proposed model - may be significant in some video surveillance applications, which control the movements and activities only in specific regions (e.g. in exhibitions, or restricted zones), while the possibly frequent motion outside that are is irrelevant.

Proposed method: The proposed method has three main parameters which we evaluated:

- $\bar{f}$ defines the minimum number of pixels under the sector required for maximal output, thus it controls the dynamic range of the feature (see Sec. 3.2);
- d_0 defines the minimal feature value required for object acceptance (see (11) and Fig. 6), we also use the notation $D_0 = 1/d_0$;
- R is the radius of the cylinders representing people in the object model (see Sec. 3.3).

Thus our evaluation is limited to these parameters only, and the remaining parameters are set as follows. In the feature extraction step (Sec. 3.2) we assumed that the sector radius was set to $r = 25$cm, the angle range Δ was to constant 30°, and the penalty parameter to $\alpha = 1.0$. In the quality function (11) D was set to constant 8. As for the parameters of the Multiple Birth and Death optimization process, we followed the guidelines provided in [3], 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.

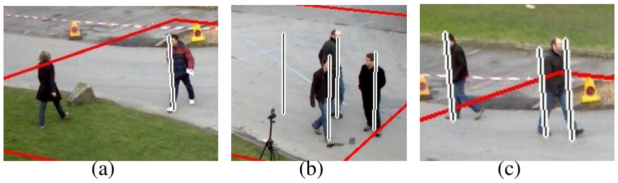


Figure 8. Examples of false localization results. (a) missed detection: the left person was not detected. (b) false detection: at the position of detection on the left no person is present. (c) multiple instances: the person on the right was localized multiple times.

Table 1. Numerical evaluation of POM on 395 frames and 1554 objects for different $\mathbf{T}$ threshold values; and comparative results of the proposed model (3DMPP) with optimized parameters $D_0 = 3.0$, $\bar{f} = 0.6$, $R = 45$.

	POM				3DMPP
T	0.850	0.875	0.900	0.950	-
MDR	2.25	2.70	3.28	4.44	4.50
FDR	15.51	15.19	14.93	14.03	2.44
MIR	0.06	0.13	0.00	0.00	0.96
TER	**17.83**	18.02	18.21	18.47	**7.91**
Recall	97.75	97.30	96.72	95.56	95.5
Prec I	86.31	86.50	86.63	87.20	97.5
Prec II	86.26	86.40	86.63	87.20	96.5

4.1. Numerical Comparison

The false localization results (MD, FD, MI) are expressed in percent of the number of all objects, we denote these ratios by MDR, FDR and MIR. Additionally, we also calculated the total false detection rate TFDR = FDR + MIR, the total error rate TER = MDR + TFDR, the recall, and the precision (Prec I is calculated from FDR, and Prec II from TFDR).

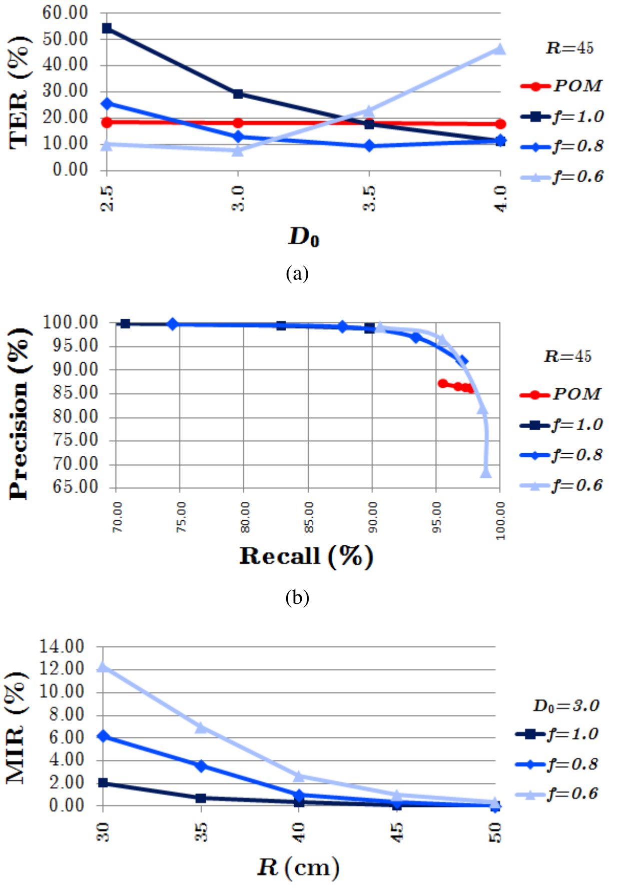
The evaluation results obtained by POM with different $\mathbf{T}$ thresholds are shown in columns 2-5 of Table 1.

Regarding the proposed model, Fig. 9(a) shows the TER plots as a function of the d_0 and $\bar{f}$ parameters, and Fig. 9(b) the Precision/Recall curves for different $\bar{f}$ settings. We have observed that among the error rates, the MIR factor depends mostly on the R parameter. Thus we demonstrate MIR curves with different $\bar{f}$ and R values in Fig. 9(c).

For easier comparison, we have also given the observed optimal 3DMPP evaluation rates in the last column of Table 1. Considering the Total Error Rate (TER) we can observe a nearly 10% gain versus the best POM result. The optimal POM rates can be also followed in Fig. 9(a) and 9(b).

5. Conclusion

In this paper we presented a novel method to localize people in multiple cameras. For this tasks we extract a pixel-level feature, which is based on the physical properties of the 2-D image formation, and produces high response for the real position and height of a person. To get a robust tool for cluttered scenes with high occlusion rate, our approach fuses features 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 [14] to extract foreground. For evaluation we used the images of a public outdoor dataset, containing three camera views and compared our method to a state-of-the-art method (POM). According to our tests, the proposed method produces accu-



(a)

(b)

(c)

Figure 9. Evaluation of the proposed 3DMPP model with different parameter settings. (a) TER as a function of the d_0 and $\bar{f}$ parameters. Optimal POM performance is demonstrated by red for comparison. (b) Prec/Recall curves in function of $\bar{f}$. (c) MIR as a function of R and $\bar{f}$ parameters.

rate estimation, even in cluttered environment, where full or partial occlusion is present and achieves significantly lower error rate than POM, especially in cases when foreground masks of people outside the area of interest overlap with the monitored area. In the future we will examine the advance of using the optimization result in the estimation process of the subsequent time step. Another possible improvement might be the use of a robust body part detector (*e.g.* [16]) in the features extraction. This can be easily integrated in the proposed algorithm with minimal modification.

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. The work of the second author was also supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. We are grateful for the authors of the reference method [4], especially for Jérôme Berclaz for his constant help in the evaluation.

References

[1] Cs. Benedek and T. Szirányi. Bayesian foreground and shadow detection in uncertain frame rate surveillance videos. *IEEE Trans. on Image Processing*, 17(4):608–621, 2008.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39(1):1–38, 1977.

[3] X. Descombes, R. Minlos, and E. Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *J. of Math. Imaging and Vision*, 33(3):347–359, 2009.

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):267–282, 2008.

[5] A. Gamal Eldin, X. Descombes, and J. Zerubia. Multiple birth and cut algorithm for point process optimization. In *Proc. of IEEE SITIS*, pages 35–42, 2010.

[6] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *Proc. of CVPR*, pages 2913–2920, 2009.

[7] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *Proc. of ECCV*, pages 324–337, 2010.

[8] L. Havasi and Z. Szlávik. 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*, pages 1275–1281, 2009.

[9] J. Kang, I. Cohen, and G. Medioni. Tracking people in crowded scenes across multiple cameras. In *Proc. of ACCV*, 2004.

[10] S. M. Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(3):505–519, 2009.

[11] I. Mikic, S. Santini, and R. Jain. Video processing and integration from multiple cameras. In *Proc. of the Image Understanding Workshop*, pages 183–187, 1998.

[12] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. *Int. J. of Computer Vision*, 51(3):189–203, 2002.

[13] PETS. Performance Evaluation of Tracking and Surveillance, 2009. http://www.cvg.rdg.ac.uk/PETS2009/a.html.

[14] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):747–757, 2000.

[15] R. Y. Tsai. 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(4):323–344, 1987.

[16] B. Wu and R. Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *Int. J. of Computer Vision*, 82(2):185–204, 2009.

[17] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13, 2006.