A Bayesian Approach on People Localization in Multi-Camera Systems

Ákos Utasi and Csaba Benedek

***Abstract*—In this paper we introduce a Bayesian approach on multiple people localization in multi-camera systems. First, pixel-level features are extracted, which are based on physical properties of the 2-D image formation process, and provide information about the head and leg positions of the pedestrians, distinguishing standing and walking people, respectively. Then features from the multiple camera views are fused to create evidence for the location and height of people in the ground plane. This evidence accurately estimates the leg position even if either the area of interest is only a part of the scene, or the overlap ratio of the silhouettes from irrelevant outside motions with the monitored area is significant. Using this information we create a 3-D object configuration model in the real world. We also utilize a prior geometrical constraint, which describes the possible interactions between two pedestrians. To approximate the position of the people, we use a population of 3-D cylinder objects, which is realized by a Marked Point Process. The final configuration results are obtained by an iterative stochastic energy optimization algorithm. The proposed approach is evaluated on two publicly available datasets, and compared to a recent state-of-the-art technique. To obtain relevant quantitative test results, a 3-D Ground Truth annotation of the real pedestrian locations is prepared, while two different error metrics and various parameter settings are proposed and evaluated, showing the advantages of our proposed model.**

***Index Terms*—Multi-camera people detection, Marked Point Process.**

I. Introduction

DETECTING and localizing people are key issues in many surveillance applications, e.g. person tracking, people counting, crowd analysis, event detection, etc. The task is still challenging in cluttered, crowded or outdoor scenes due to the high occlusion rate between the different moving and static scene objects. By applying background subtraction 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 break apart [1]–[3]. Under such conditions single-camera localization approaches might be inefficient: a straightforward improvement is to utilize images of different cameras from different viewpoints simultaneously. 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 can 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 to the ground and on multiple parallel planes. Our approach does not use any color or shape models for distinguishing the people in the scene. Instead, we exploit the advantage of multiple cameras, and from the projected foreground masks two similar pixel-level features are extracted in each 2-D top-view position: one on the ground plane, and one on the estimated head plane. Both features collect evidence for the existence of a person at a given position. In addition, we distinguish two different gait phases and derive separate descriptors to indicate pedestrians in stance and in swing phase [4], respectively. Finally, the extracted features are used in a stochastic optimization process with geometric constraints to find the optimal configuration of multiple people, as partially introduced in [5], [6].

The rest of the paper is organized as follows. In Sec. II we briefly present the related work in single- and multi-camera people detection. The proposed method is discussed in Sec. III. In Sec. IV we evaluate our method using two public datasets.

II. Related work

In the last decades single-camera person detection has undergone a great evolution. See [7] for an extensive review of State-of-the-Art (SoA) 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-camera approaches provide a better solution, that can accurately estimate the position of multiple people. Mikic *et al.* [8] proposed a blob based approach, where one object is represented by one blob on each view, and 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. [9] models the appearances (color) and locations of the pedestrians, to segment people on camera views. This step helps the separation of foreground regions belonging to different objects. [10] extracts moving foreground blobs, and calculates the centroid of the blob's lowest pixels, which is projected to 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 are inefficient in cluttered environment where the objects break apart or when the occlusion rate is high. Therefore, the pixel-level constraints applied in our model only concern specific object parts.

The method called Probabilistic Occupancy Map (POM) [11] assumes that the objects are observed by multiple cameras

The authors are with the Distributed Events Analysis Research Laboratory, Computer and Automation Research Institute of the Hungarian Academy of Sciences, Kende u. 13-17, H-1111 Budapest, Hungary. E-mail: {utasi,bcsaba}@sztaki.hu

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. According to the authors, the method is mainly affected by the incorrect foreground blobs obtained by background subtraction (*e.g.* reflections, or cast shadow). In contrast, as shown later in our evaluation, the feature extraction step of our method significantly reduces these errors, and eliminates false detections originating from the foreground blobs of people outside the monitored area. The method in [12] fuses evidence from multiple cameras to find image locations of scene points that are occupied by people. The homographic occupancy constraint is proposed, which fuses foreground likelihood information from the camera images 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 that 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 evidence 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. Although camera calibration is a prerequisite of our method, which is performed manually in most cases, we do not think this assumption is very restrictive since it should be done only once and minor displacements can be easily re-corrected using existing methods (*e.g.* [13]).

State-of-the-Art methods can also be grouped based on the approach of object modeling. *Direct* techniques construct the objects from primitives, like silhouette blobs [14] 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* assign a fitness value to each possible object configuration and an optimization process attempts to find the configuration with the highest confidence. Marked Point Processes (MPP) [15] provide efficient tools to extend the well established Markov Random Field (MRF) [16] based pixel level classification techniques, by taking into account the geometry in the proposed models. In an MPP model, the variables are the objects instead of the pixels, and populations of unknown number of objects can be jointly handled in a Baysian framework. Moreover, similarly to MRFs, MPPs can also embed prior constraints and data models within a global configuration probability function, and various techniques for optimizing the models [15], [17] and estimating the parameters [18] are available.

In [19] a single view MPP model is developed to detect and count people in crowded scenes. The model couples a spatial stochastic process governing the number and placement of individuals with a conditional mark process for selecting body shape. However, limitations of the monocular approach create difficulties in strongly crowded scenarios, where the overlap rate is high. Thus in the presented method, we optimize the objects in the 3-D real world space instead of the 2-D shapes in the individual camera views, similarly to [20]. The main difference from the latter approach lies in data model construction, as our proposed pixel-level feature focuses on the accurate extraction of the foot and head positions 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, meanwhile 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 that tends to be sensitive to false local maxima resulting ghost effects [20], we apply the recently proposed Multiple Birth-and-Death (MBD) technique [17] that by design is 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 of adding several random objects to the current configuration. In addition, there is no rejection during the birth move, therefore high energy objects can still be added independently of the temperature parameter - this property prevents the algorithm from being stuck at ghost objects.

The paper's main contributions are the following. Firstly, we developed a new method for multi-view people detection and localization. The method extracts a set of novel pixel-level features from the foreground pixels projected on multiple parallel planes. Thereby, we avoid the usage of unreliable object-level features like color or shape. The extracted features are embedded into a MPP framework, and the final configuration, *i.e.* locations of the people, are obtained by an efficient optimization technique. Secondly, we defined two different error metrics for the numerical evaluation of the localization accuracy, which allow other methods to be compared against our results whether they estimate the 3-D world ground coordinates of people or the 2-D positions on the camera images. We manually annotated two public datasets, and performed numerical evaluation to demonstrate the accuracy of our approach. We compared our method to a State-of-the-Art technique, and according to our tests the proposed method achieves superior performance in most cases.

III. Proposed method

The input of the proposed method consists of foreground masks extracted from multiple calibrated camera views [21], monitoring the same scene. In our current implementation the masks are obtained 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, observable from a virtual birds-eye viewpoint above the ground plane. However, no prior information of a person's height is known, and the height of different people in the scene

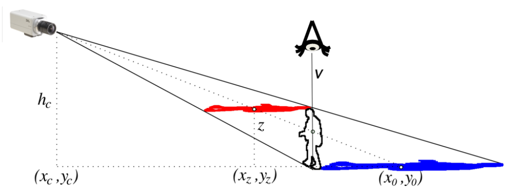


Fig. 1. Silhouettes are projected to the ground plane (blue) and to parallel planes (red).

may also be different. Therefore, we project the silhouette masks on multiple parallel planes at heights in the range of typical human height. In crowded scenes the overlap rate is usually high, which could 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 following subsections:

1) *Multi-plane projection:* The silhouettes are projected to the ground and to several parallel planes at different heights.
2) *Feature extraction:* At each location of each plane we extract pixel-level features that provide a 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.

A. Multi-plane projection

Let us denote the ground plane with P_0, and by P_h a parallel plane above P_0 at height h. In the first step of the proposed method we project the detected silhouettes to P_0 and to different P_h planes (with different $h > 0$ offsets) by using the model of the calibrated cameras. As shown Fig. 1, this can be efficiently performed by projecting onto P_0 only, then using the following relationship. Let (x_c, y_c) denote the ground 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 P_0. Then the (x_h, y_h) position of the same point projected to a parallel plane at h height can be expressed as

$$\begin{aligned} x_h &= x_0 - (x_0 - x_c)\, h/h_c \\ y_h &= y_0 - (y_0 - y_c)\, h/h_c \end{aligned} \tag{1}$$

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

B. 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 in the conventional pinhole camera model. Let us consider in Fig. 1 the person with *real* height $h^\star$, and project the silhouette onto the P_0 ground plane (marked with blue) and on the P_h plane at the *estimated* height of the person (*i.e.* $h = h^\star$, 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 $P_{h|h=h^\star}$ 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_h, looking down on the ground plane in a perpendicular direction. Here, the silhouette prints from P_h and P_0 are projected to a common P_{x-y} plane and jointly shown by red and blue colors respectively, and overlapping areas are purple. We can observe in Fig. 2(a), that if the height estimation is correct (*i.e.* $h = h^\star$), the two prints just touch each other in the $\mathbf{p} = (x, y)$ point, which corresponds to the ground position estimate of the person. However, if the person's height is underestimated (*i.e.* $h < h^\star$), the two silhouette prints will overlap as shown in Fig. 2(b), and when the height is overestimated (*i.e.* $h > h^\star$), the silhouettes will move away, see Fig. 2(c).

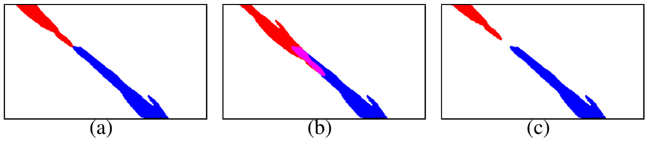


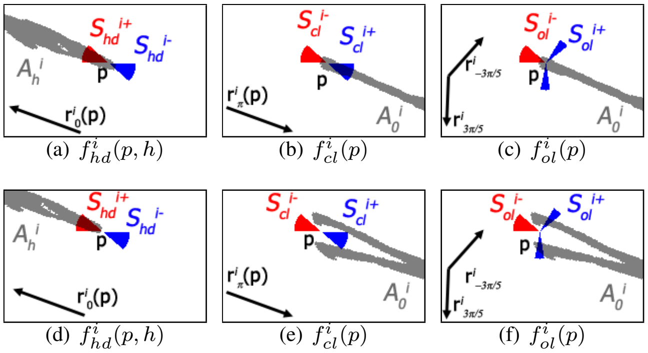
Fig. 2. Our features are based on the 2-D image formation properties and on the multi-plane projection representation. The ground plane projection of one silhouette is marked with blue, and the P_h plane projection for three different h values (h is the distance from the ground) with red. (a) projection for h equals to the person's real height; (b) projection for h lower than the person's real height; (c) projection for h higher than the person's real height.

The next task is to define numerical features which evaluate a given $[\mathbf{p}, h]$ object candidate - with estimated ground position $\mathbf{p}$ and height h - based on the multi-camera information. Since mapping the silhouettes from the camera views to the joint top view is nonlinear, we need to find descriptors which are insensitive to the geometric distortions of the body shapes. We denote by $\mathbf{r}_0^i(\mathbf{p})$ a unity vector, which points from $\mathbf{p}$ towards the vertical ground projection of the ith camera onto the P_0 plane, and by $\mathbf{r}_\varphi^i(\mathbf{p})$ the rotation of $\mathbf{r}_0^i(\mathbf{p})$ with angle φ around the vertical axis v. We denote the foreground points projected to the P_0 and P_h planes by A_0^i (blue regions in Fig. 1 and Fig. 2) and A_h^i (red regions), respectively.

Based on the above observations, an object hypothesis $[\mathbf{p}, h]$ is relevant according to the ith camera data if it jointly meets constraints about the *head* and *leg* positions. *On one hand*, we should find projected silhouette pixels on the P_h head plane (*i.e.* red prints) in the neighborhood of the $\mathbf{p}$ point in the $\mathbf{r}_0^i(\mathbf{p})$ direction, but penalize such silhouette points in the opposite direction $\mathbf{r}_\pi^i(\mathbf{p})$. To measure this property, we define circular *head* (*hd*) sectors $S_{hd}^{i+}(\mathbf{p})$ and $S_{hd}^{i-}(\mathbf{p})$ around $\mathbf{p}$ directed into $\mathbf{r}_0^i(\mathbf{p})$ (red in Fig. 3) and $\mathbf{r}_\pi^i(\mathbf{p})$ respectively. The sectors have fixed arc and radius, being the parameters of the model. Then, following Fig. 3(a) and (d), we calculate the $f_{hd}^i(\mathbf{p}, h)$ *head* feature at height h as:

$$f_{hd}^i(\mathbf{p},h) = \frac{\mathbf{Area}\big(A_h^i \cap S_{hd}^{i+}(\mathbf{p})\big) - \mathbf{Area}\big(A_h^i \cap S_{hd}^{i-}(\mathbf{p})\big)}{\mathbf{Area}\big(S_{hd}^{i+}(\mathbf{p})\big)}\ . \tag{2}$$

On the other hand, we distinguish two different cases by the definition of the *leg* position constraint. People with *closed legs* (*cl*, standing, or in the stance phase of the gait cycle) can



(a) $f_{hd}^i(p,h)$ (b) $f_{cl}^i(p)$ (c) $f_{ol}^i(p)$

(d) $f_{hd}^i(p,h)$ (e) $f_{cl}^i(p)$ (f) $f_{ol}^i(p)$

Fig. 3. Calculation of the $f_{hd}^i(p,h)$, $f_{cl}^i(p)$ and $f_{ol}^i(p)$ features in two selected positions, corresponding to a person with closed (top) and open (bottom) legs, respectively.

be handled in an analogous manner to the *head* feature (see Fig. 3(b)). Here S_{cl}^{i+} and S_{cl}^{i-} sectors correspond to $\mathbf{r}_\pi^i(\mathbf{p})$ and $\mathbf{r}_0^i(\mathbf{p})$ directions respectively, and the feature is:

$$f_{cl}^i(\mathbf{p}) = \frac{\mathbf{Area}\big(A_0^i \cap S_{cl}^{i+}(\mathbf{p})\big) - \mathbf{Area}\big(A_0^i \cap S_{cl}^{i-}(\mathbf{p})\big)}{\mathbf{Area}\big(S_{cl}^{i+}(\mathbf{p})\big)}\ . \qquad (3)$$

However, if the person is in the swing phase of the gait cycle the previous descriptor proves to be inaccurate (see Fig. 3(e)). Therefore, we have developed an *open leg* (*ol*) feature (see Fig. 3(c) and 3(f)), whose attractive region, S_{ol}^{i+}, consists of two, half sized circular sectors corresponding to the directions $\mathbf{r}_{\pm 3\pi/5}^i(\mathbf{p})$. The repulsive sector, S_{ol}^{i-} is constructed in the same way as S_{cl}^{i-}. Then, the $f_{ol}^i(\mathbf{p})$ feature term is derived similarly to $f_{cl}^i(\mathbf{p})$. Since we have observed that for our purposes, the gait phase of each person can be fairly approximated either by the *closed* or by the *open* leg states, the *joint leg* feature $f_l^i(\mathbf{p})$ is obtained as

$$f_l^i(\mathbf{p}) = \max\left[f_{cl}^i(\mathbf{p}), f_{ol}^i(\mathbf{p})\right]\ . \qquad (4)$$

Note that while the *leg* feature $f_l^i(\mathbf{p})$ is calculated only from the ground plane P_0 projections, the value of the *head* feature $f_{hd}^i(\mathbf{p},h)$ depends on the height h of the plane, where the foreground is projected. In our method h takes values in the range of typical human heights, *i.e.* P_h estimates the *head* level of the person.

Finally, the *head* and *leg* features are truncated to take values in the $[0,\hat{f}]$ range, and are normalized by $\hat{f}$. Here, $\hat{f}$ controls the area ratio required to produce the maximal output, *i.e.* it is the dynamic range parameter of the feature. Further below $f_{hd}^i(\mathbf{p},h)$ and $f_l^i(\mathbf{p})$ refer to this normalized value.

If the object defined by the $[\mathbf{p},h]$ parameters is completely visible for the ith camera, both the $f_{hd}^i(\mathbf{p},h)$ and $f_l^i(\mathbf{p})$ features should have *high* values. However, in the available views, some of the legs or heads may be partially or completely occluded by other pedestrians or static scene objects, which can strongly corrupt the feature values. Although the descriptors may be weak in the individual cameras, we can construct a stronger feature if we average the responses of the N available cameras, *i.e.*

$$\bar{f}_{hd}(\mathbf{p},h) = \frac{1}{N}\cdot\sum_{i=1}^{N} f_{hd}^i(\mathbf{p},h)\ ,\quad \bar{f}_l(\mathbf{p}) = \frac{1}{N}\cdot\sum_{i=1}^{N} f_l^i(\mathbf{p})\ . \qquad (5)$$

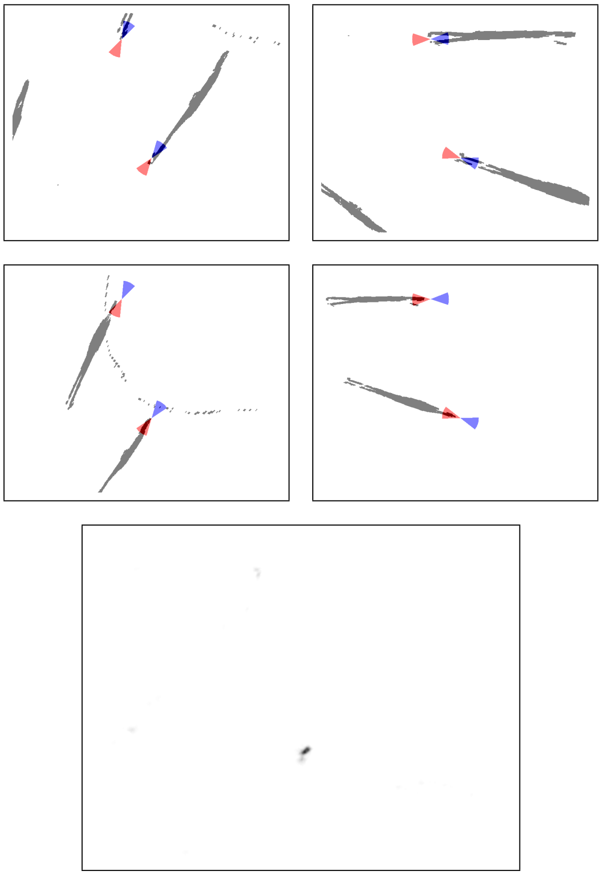


Fig. 4. Top: $f_{cl}^i(\mathbf{p})$ closed leg feature extraction from two camera views; Center: $f_{hd}^i(\mathbf{p},168\text{cm})$ head feature extraction from two camera views on plane $P_{h=168\text{cm}}$. The sector pointing in the camera's direction $\mathbf{r}_0^i(\mathbf{p})$ is denoted by red, in the opposite direction $\mathbf{r}_\pi^i(\mathbf{p})$ by blue color. Bottom: $f(\mathbf{p},168\text{cm})$ strong features calculated by fusing the features of the top and center figures. Lower intensity pixels indicate more probable person positions.

Finally, the joint data feature $f(\mathbf{p},h)$ is derived as

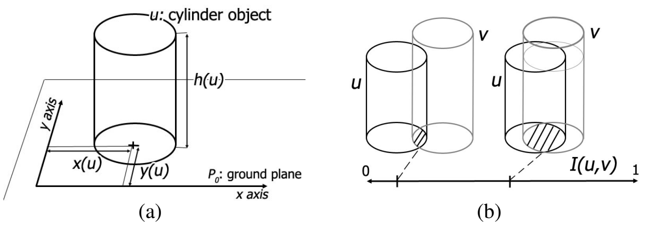
$$f(\mathbf{p},h) = \sqrt{\bar{f}_{hd}(\mathbf{p},h)\cdot\bar{f}_l(\mathbf{p})}\ , \qquad (6)$$

to express that high *head* and *leg* feature responses should be *jointly* present in the same 3-D position assuming a given height. Fig. 4 demonstrates the joint feature $f(\mathbf{p},h)$ using two camera views and assuming $h = 168$cm. Lower intensity pixels are used to show more probable $\mathbf{p}$ positions for the given h. We can observe that the height of the bottom person is close to the estimated 168cm, since the fused feature indicates a high probability value at the ground position of the person.

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.

C. 3-D Marked Point Process model

In this section, we propose a 3-D Marked Point Process model (3DMPP) to detect and localize the people in the scene, and provide their position and height parameters. The Marked



(a) (b)

Fig. 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.

Point Process (MPP) framework enables to characterize complete object populations instead of individual persons, through exploiting information from entity interactions. MPP techniques follow a Bayesian approach: first a probability density function is defined on the configuration space, thereafter the population with the highest probability is estimated by an optimization process [22]. Following the classical Markovian approach, each object may only affect its *neighbors* directly. This property keeps the number of interactions in the population tractable, and results in a compact description of the global scenario that can be analyzed with efficient algorithms.

We approximate a person by a cylinder u in the 3-D coordinate system of the scene, with a fixed radius R. Let us assume that the ground is flat and the people are standing on it. We monitor a rectangular Area of Interest (AOI) in the P_0 ground plane, and we attempt to detect each pedestrian, whose foot is inside this AOI. (However, the silhouettes in the individual camera views may partially fall outside the projected AOI regions.) Thus the free parameters of a given object-cylinder u are its $\mathbf{p} = (x, y)$ coordinates in the ground plane and the h height of the cylinder, as shown in Fig. 5(a).

In the implementation, we use a discrete space of the objects: we discretize the AOI 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 an arbitrary number of cylinder objects in the scene. Thus the Ω configuration space is defined as:

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \qquad \Omega_n = \Big\{ \{u_1, \dots, u_n\} \in \mathcal{H}^n \Big\} \ . \tag{7}$$

Let ω denote an arbitrary object configuration $\{u_1, \dots, u_n\}$. To enable considering the Markov-property of the configuration model, we first define a $\sim$ neighborhood relation between the objects in $\mathcal{H}$. In our model, the $u \sim v$ relation holds if the cylinders of u and v intersect.

We refer to the global input data as $\mathcal{D}$, in the model, which consists of the foreground silhouettes in all camera images and the camera calibration matrices. For characterizing a given ω object population considering $\mathcal{D}$, we introduce a non-homogeneous, data-dependent Gibbs distribution on the configuration space:

$$P_{\mathcal{D}}(\omega) = \frac{1}{Z} \cdot \exp\Big[-\Phi_{\mathcal{D}}(\omega) \Big] \ , \tag{8}$$

where Z is a normalizing constant: $Z = \sum_{\omega \in \Omega} \exp\big[-\Phi_{\mathcal{D}}(\omega) \big]$, and $\Phi_{\mathcal{D}}(\omega)$ is the configuration energy function, 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:

$$\Phi_{\mathcal{D}}(\omega) = \sum_{u \in \omega} J_{\mathcal{D}}(u) + \gamma \cdot \sum_{\substack{u, v \in \omega \\ u \sim v}} I(u, v) \ , \tag{9}$$

where $J_{\mathcal{D}}(u) \in [-1, 1]$, $I\,(u, v) \in [0, 1]$ and γ is a weighting factor. Note that the role of $I\,(u, v)$ is similar to the *smoothness term* in MRF models. We derive the optimal object configuration as the maximum likelihood configuration estimate, obtained as

$$\omega_{\text{ML}} = \underset{\omega \in \Omega}{\operatorname{argmin}} \big[\Phi_{\mathcal{D}}(\omega) \big] \ . \tag{10}$$

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 containing 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)) :

$$I(u, v) = \frac{\mathbf{Volume}\big(u \cap v\big)}{\mathbf{Volume}\big(u \cup v\big)} \ . \tag{11}$$

Without the above geometrical constraint we could easily place many cylinder objects into the most probable person positions of the Fig. 4, where the low intensity pixels denote these positions. Since $I(u, v)$ penalizes nearby cylinder objects, it will efficiently select only one cylinder for one person.

On the other hand, the $J_{\mathcal{D}}(u)$ *unary* potential characterizes a proposed object candidate segment $u = [\mathbf{p}, h]$ depending on the multi-camera image data, but independent of other objects of the population. Cylinders with negative unary potentials are called *attractive objects*. Considering (9) 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 $f_u = f(\mathbf{p}, h)$ feature at ground point $\mathbf{p}$ in the 3DMPP model, introduced in Sec. III-B. Let us keep in mind, 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):

$$J_{\mathcal{D}}(u) = Q(f_u, d_0) = \begin{cases} \left(1 - \frac{f_u}{d_0}\right) & \text{if } f_u < d_0 \\ \exp\left(-\frac{f_u - d_0}{8}\right) - 1 & \text{if } f_u \geq d_0 \end{cases} \tag{12}$$

where d_0 is parameter, and the denominator 8 performs data-normalization. Consequently, object u is attractive according to the $J_{\mathcal{D}}(u)$ term iff $f_u \geq d_0$. Thus the d_0 parameter defines the minimal feature value required for object acceptance. In

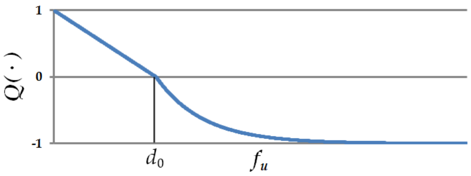


Fig. 6. Plot of the $Q(f_u, d_0)$ function (12) used for projecting the feature domain. The d_0 parameter defines the minimal value for object acceptance.

the subsequent sections we also use the notation $D = 1/d_0$ as a shorthand.

D. Optimization by multiple birth-and-death dynamics

Even with the discretization of the $\mathbf{p} = (x, y)$ and h object descriptors, and prescribing at most one person in a given ground position as constraint, the cardinality of the population space is an exponential function of the number of possible locations. For example, for our PETS data we used 609×745 locations in P_0 and 55 different height values (between 155 and 210 cm, with 1cm accuracy), yielding $(55+1)^{609\times745}$ different configurations - as each location may be empty, or contain a person with arbitrary height. Thus exhaustive search cannot be used for minimizing (10), instead, we should adopt techniques that can efficiently sample the configuration space.

In previous MPP applications, various optimization methods have been utilized [23], mainly implementing an iterative process consisting of object proposition (birth) and removal (death) steps. The most widely used approach has been the RJMCMC technique [20], [24], where in the birth step, moves are added such as split, translate, rotate, etc. The main limitation here is that each iteration consists of perturbing one or a couple of objects and the rejection rate induces a huge computation time. A faster algorithm - called Multiple Birth and Death (MBD) - has been proposed in [17] and adopted in [22], enabling multiple perturbations in parallel, resulting in increased speed of convergence and simplicity of implementation. For choosing a trade-off between speed and quality we have also adapted the MBD optimization to our MPP model. The steps are as follows:

Initialization: start with an empty population $\omega = \emptyset$, and let s denote a pixel of the discretized AOI (*i.e.* a 2-D pixel lattice of size $S_W \times S_H$).

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

Let us observe that the number of objects, n, is not an external parameter of the model, but it evolves during the iteration steps. It mainly depends on the input feature maps, however, the density of objects is also influenced by the non-overlapping term of the configuration energy.

IV. Experiments

We have compared our approach to the Probabilistic Occupancy Map (POM) technique [11], which is a state-of-the-art method with similar purposes[1]. In the POM approach, the area of interest is divided into discrete locations by a regular rectangular grid whose resolution, ν, is a parameter of the process. Then the 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 results of background-subtraction from different viewpoints. The appearance model is parametrized by a family of rectangles approximating the silhouettes of average sized individuals (with height 175cm and width 50cm) standing at every location of interest, from every point of view. The POM method outputs grid position occupancy probabilities, and then people locations are obtained by thresholding this probability map. For the detection, we use here a fixed threshold parameter, τ.

For the evaluation of the two methods we used two public sequences. First, from the PETS 2009 dataset [25] we selected the *City center* images containing approximately 1 minute of recordings (400 frames total) in an outdoor environment. From the available views we selected cameras with large fields of view (View_001, View_002, and View_003) and we used an AOI of size 12.2m × 14.9m, which is visible from all three cameras. The maximum number of pedestrians at the same time inside the AOI is 8. Fig. 7(a) shows an example frame taken from the 1st camera view, ground AOI is represented by a black rectangle with gray outline.

The second dataset we used in our experiments is the EPFL *Terrace* dataset, which is 3 minutes and 20 seconds long (5000 frames total). The scene is semi-outdoor, since it was recorded in a controlled outdoor environment and it also lacks some important properties of a typical outdoor scene (*e.g.* no background motion caused by the moving vegetation is present, and no static background objects occlude some parts of the scene). We selected three cameras having small fields

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

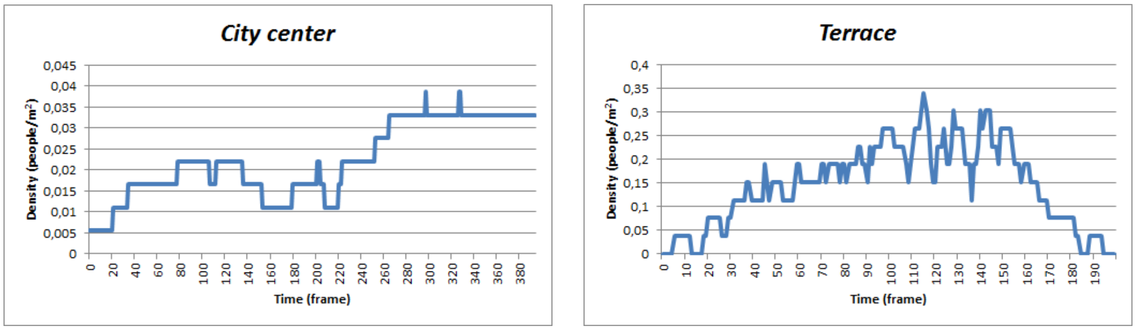
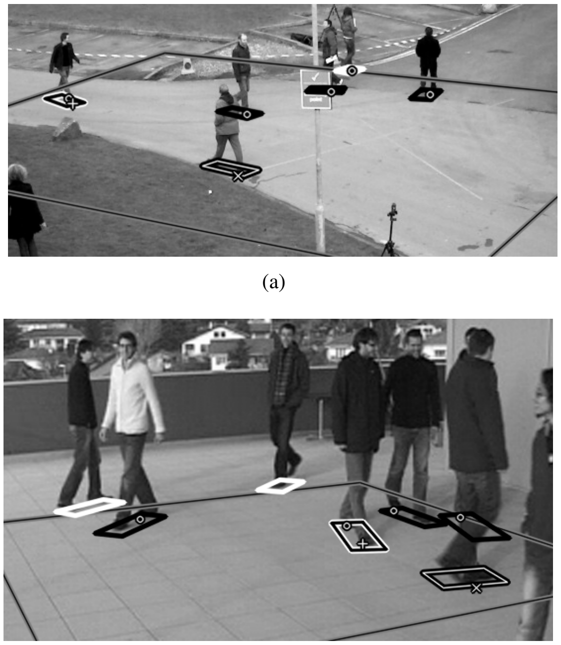


Fig. 8. Comparison of the used datasets: people density over time in the *City center* (left), and in the *Terrace* (right) sequences. The two sequences have different dynamics, a more severe crowd and a higher occlusion rate are present in the *Terrace* sequence.



(a)

(b)

Fig. 7. Example frames from the (a) *City center* and (b) *Terrace* sequences. Rectangles represent the ground truth annotation, the $\circ$, $+$, and $\times$ symbols denote the estimated ground positions. Bold black rectangles with $\circ$ symbol denote the successful detections (*e.g.* the rightmost person of (a)), white rectangles with black outline represent missed detections (MD, *e.g.* the most bottom person in (b)), black rectangles with white outline and a $+$ symbol show the multiple instances (MI, *e.g.* the leftmost person of (a)), and the $\times$ symbols denote the false detections (FD). Ground truth rectangles partially inside the AOI are denoted by bold white rectangles (*e.g.* the leftmost person in (b)). The AOI is represented by a black rectangle with gray outline.

of view, and defined the AOI as a $5.3\text{m} \times 5.0\text{m}$ rectangle. The scene is severely cluttered in some periods, the maximum number of pedestrians at the same time inside the AOI is 8, see Fig. 7(b). Also note that the *City center* and *Terrace* datasets have different characteristics with respect to the density of people inside the AOI, as shown in Fig 8. It can be clearly seen that the *Terrace* sequence contains more severe crowd, and thereby a higher occlusion rate.

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 [26], 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 [1] was used with the following settings: modality parameter $T_{\text{bg}} = 0.6$, matching criterion $I_{\text{bg}} = 3.0$. During the evaluation of POM, we manually masked out the regions on each camera that do not belong to the volume of interest defined by a rectangular cuboid[2]. Our method does not require such region masking, therefore this step was neglected in the evaluation of our method.

While the above method produced fairly accurate foreground masks for both sequences, there are some practical problems that should be considered. First, the foreground silhouette of people might brake apart or partly disappear mainly due to other static objects in the scene (*e.g.* trees, traffic signs, or electric wires, see Fig. 9(a)) or get merged with the masks of other objects when they overlap each other as shown in Fig. 9(b). Fig. 9(c) shows an example when these problems occur at the same time in all camera views. Moreover, most cameras have a built-in automatic mechanism to shift the white balance when the scene changes significantly. When the fields of view is small, large objects close to the camera can trigger this camera function. Fig. 9(d) demonstrates this problem, where the dark clothes of several people significantly change the characteristic of the scene. Nevertheless, our proposed method detects correctly 5 out of the 6 pedestrians of Fig. 9(d), even the two ones standing partially inside the AOI (see white rectangles).

A. Evaluation methodology

For numerical evaluation we created 3-D ground truth annotations for both the *City center* and *Terrace* multi-camera sequences as follows. The ground occupancy of each person is represented by a rectangle on the ground plane covering the area of the human body between the two leg positions. The center position, size and orientation of this rectangle is estimated manually by projecting the rectangle to each camera view. Finally, we increased the areas of rectangles when the projections had significant difference (caused *e.g.* by synchronization error or calibration inaccuracy, see Fig. 10). Thus our annotation at a given time stamp is an $\mathbf{R} = \{r_1, \ldots, r_m\}$ set of m rectangles on the ground plane, where each r_i

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

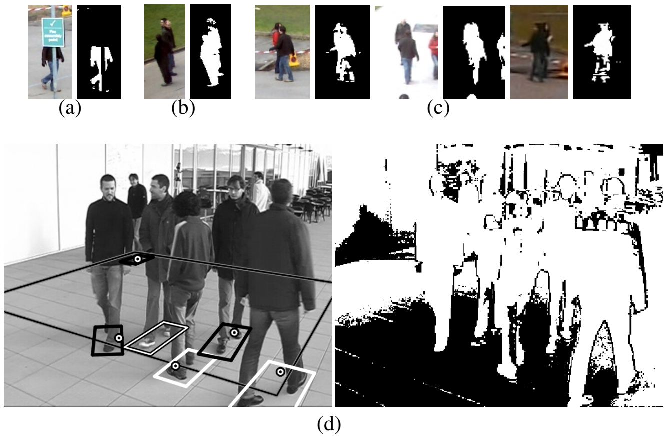


Fig. 9. Practical problems of foreground-background separation: foreground masks often (a) brake apart or (b) get merged with the masks of other objects, (c) these problems might occur at the same time in all camera views. (d) Objects close to the camera might trigger the automatic white balance function of the camera. In this particular example it was caused by the dark clothes of the people entering the fields of view of the camera. Background is denoted by black, and foreground by white color.

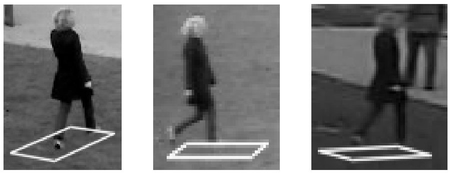


Fig. 10. Inaccurate camera calibration and synchronization error: the white ground truth rectangle is projected to all three views.

rectangle is parameterized by its ground position $x(r_i), y(r_i)$, size $w(r_i), h(r_i)$ (width and height), and orientation $\theta(r_i)$. Fig. 7 demonstrates several ground truth rectangles, where different colors represent different error types, to be discussed later.

The two methods estimate the ground occupancy of detected people using different models. However, in both cases we can easily compute the estimated ground position of a person. In case of the proposed method we simply used the center of the cylinder model (see Sec. III-C), while for POM we took the center of the cell which was assigned to the person in the rectangular grid projected to the ground. Thus in the comparison a detected p person is represented by its ground position $x(p)$ and $y(p)$. The set of detected people at a given timestep is denoted by $\mathbf{P} = \{p_1, \ldots, p_n\}$.

Given the ground truth data $\mathbf{R}$ and the estimated positions $\mathbf{P}$ we define a match function $m(i,j)$ to indicate whether the estimated p_j is inside the annotation r_i or not, *i.e.*

$$m(i,j) = \begin{cases} 1 & \text{if } p_j \text{ is inside } r_i \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

and we use the Hungarian algorithm [27] to find the maximum matching, *i.e.* the maximum utilization of $\mathbf{M} = [m(i,j)]_{m \times n}$. We denote by $\mathbf{A} = [a(i,j)]_{m \times n}$ the assignment obtained by the algorithm, *i.e.* $a(i,j) = 1$ if p_j was assigned to r_i and 0 otherwise. Finally, we count

- Missed Detections:

$$\mathrm{MD} = \# \left\{ r_j : \sum_{i=1}^{n} a(i,j) = 0 \right\} ,$$

i.e. no estimation was assigned to the ground truth (represented by white rectangles with black outline Fig. 7);

- False Detections:

$$\mathrm{FD} = \# \left\{ p_i : \sum_{j=1}^{m} a(i,j) = 0 \right\} ,$$

i.e. no ground truth could be assigned to an estimate (represented by × symbols in Fig. 7);

- Multiple Instances:

$$\mathrm{MI} = \sum_{j=1}^{m} \max \left(0, \sum_{i=1}^{n} a(i,j) - 1 \right) ,$$

i.e. multiple estimates were assigned to a ground truth (represented by black rectangles with white outline and a + symbol in Fig. 7);

- Total Error:

$$\mathrm{TE} = \mathrm{MD} + \mathrm{FD} + \mathrm{MI} .$$

It is difficult to decide visually whether the person near the borders of the AOI should be considered as being inside or outside. Therefore, we included the following two solutions. Firstly, we created our annotations in a larger AOI using an additional 25cm buffer zone. Secondly, we neglected the MDs if the ratio of the area of r_j inside the AOI and the total area of r_j does not exceed 50% (represented by bold white rectangles in Fig. 7). The first step reduces the FDs, while the second one reduces the MDs occurring near the borders of the AOI.

We defined two different comparison metrics by determining $\mathbf{M}$ and $\mathbf{A}$ from

1) the real world ground truth annotation and position estimates: Ground Position Error (GPE) metric;
2) the projected ground truth and positions, and we selected the view where the TE is minimal: Projected Position Error (PPE) metric.

These two tests allow other methods to be compared against our results whether they estimate the real world ground position of people or the 2-D position on the camera images (*e.g.* using camera homography instead of calibration).

In case of the *City center* sequence we annotated all 400 frames, while the *Terrace* sequence has been annotated in 1Hz frequency resulting in 200 annotated frames[3].

B. Numerical comparison

After counting all the false localization results (MD, FD, MI) on all annotated frames we express them in percent of the number of all objects, we denote these ratios by MDR, FDR, MIR, and TER. Note that while MDR ≤ 1 and MIR ≤ 1 always hold, in case of many false alarms FDR (thus also TER) may exceed 1.

[3] The annotation data and the evaluation software will be available in the final version.

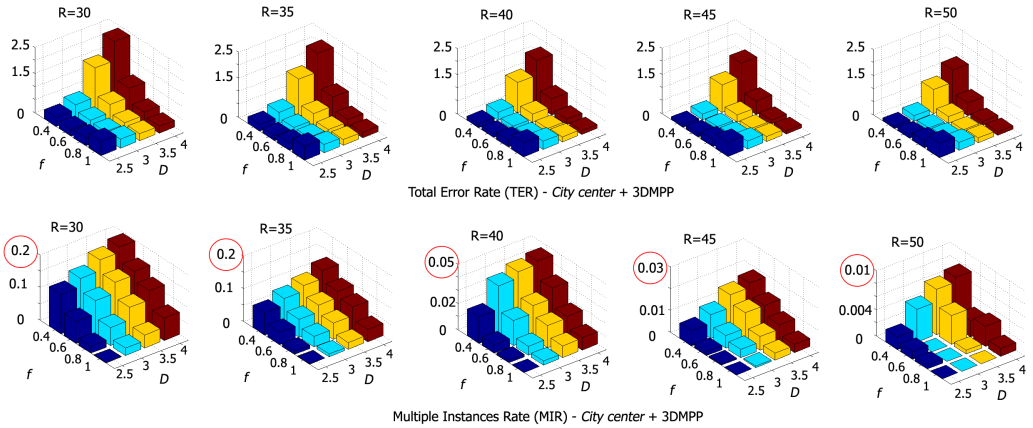


Fig. 11. Evaluation of the Total Error Rate (TER, top) and the Multiple Instances Rate (MIR, bottom) of the proposed 3DMPP model for the *City center* sequence as a function of various $\hat{f}$ and D parameter values for different R cylinder radius. Please note the marked scale differences in the bottom row graphs.

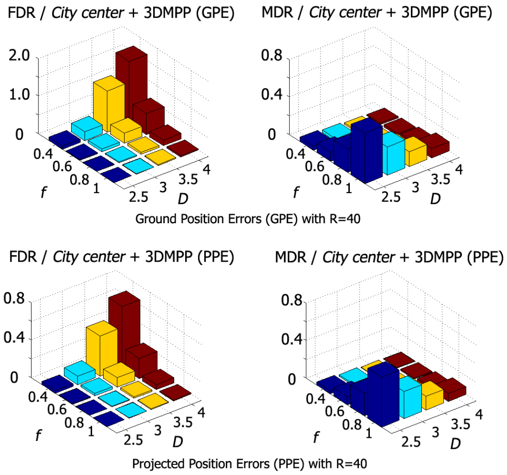


Fig. 12. Comparing the Ground Position Errors (GPE) and Projected Position Errors (PPE) of the proposed 3DMPP model for the *City center* sequence: the False Detection Rate (FDR) plots and Missed Detection Rate (MDR) plots of both metrics are shown with varying $\hat{f}$ and D parameter values and optimal $R = 40$ radius settings. Similarity of the corresponding plots confirm the appropriateness of both GPE and PPE for method comparison.

For comprehensive evaluation, we have tested the proposed 3DMPP model with various parameter settings. The proposed method has three main parameters, which we evaluated:

- $\hat{f}$ defines the minimum number of pixels under the sector required for maximal output, thus it controls the dynamic range of the feature (see Sec. III-B);
- d_0 defines the minimal feature value required for object acceptance (see (12) and Fig. 6), we also use the notation $D = 1/d_0$;
- R is the radius of the cylinders representing people in the object model (see Sec. III-C).

Thus our evaluation is limited to these parameters only, and the remaining parameters are set as follows. We used a constant 2cm grid resolution during the multi-plane projection. In the feature extraction step (Sec. III-B) we assumed that the sector radius was set to $r = 25$cm and the angle range $\varDelta$ to a constant 30°. As for the parameters of the Multiple Birth and Death

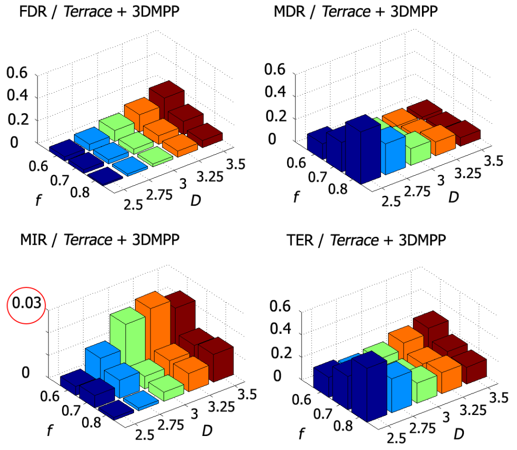


Fig. 13. Evaluating the proposed 3DMPP method's Ground Position Errors (GPE) for the *Terrace* sequence: FDR, MDR, MIR and TER plots are shown with varying $\hat{f}$ and D parameter values and a fixed radius $R = 40$. Please note the scale differences between the bottom left graph w.r.t. the others.

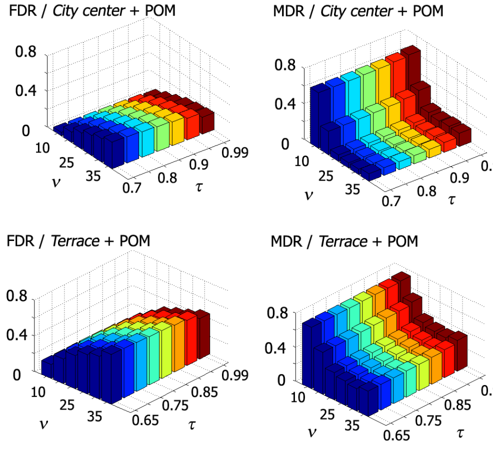


Fig. 14. Evaluating the POM method's Ground Position Errors (GPE) for the *City center* and *Terrace* sequences: the False Detection Rate (FDR) plots and Missed Detection Rate (MDR) plots of both metrics are shown with varying ν and τ parameter values.

optimization process, we followed the guidelines provided in [17], 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.

For the *City center* sequence, Fig. 11 shows the measured TER (top) and the MIR (bottom) values in the GPE metric, as a function of various $\hat{f}$ and D parameter ensembles for different R cylinder radius values. We have obtained a minimal $\mathrm{TER} = 0.123$ with parameter settings $\hat{f} = 0.8$, $D = 3$ and $R = 40$cm. We can also observe the MIR is decreased notably by increasing R from 30 to 40, since naturally, cylinder models with higher radii cause less multiple detected people. Note as well that TER's dependence on R is not significant (see Fig. 11 top), because multiple detections are in general the least frequent artifacts (*i.e.* $\mathrm{MIR} \ll \mathrm{TER}$).

To demonstrate the strong connections between the GPE and the PPE, we have displayed in Fig. 12 the FDR and MDR plots obtained by the GPE and PPE metrics with the same parameter settings. The similarity of the corresponding plots confirm that the two approaches are equivalently appropriate for method evaluation, thus PPE can be used for techniques where calibration information is not available.

By testing our model on the *Terrace* sequence, Fig. 13 shows the FDR, MDR, MIR and TER plots with $R = 40$cm radius. Here we have reached an optimum $\mathrm{TER} = 0.131$ with $\hat{f} = 0.7$ and $D = 3$.

Next, we evaluated the POM method with varying ν grid resolution and τ probability threshold parameters, the resulting FDR and MDR plots are shown in Fig. 14. We can see here that the dependence on τ is less significant, since the probability values of POM after convergence tend to be close either to 0 or to 1. On the other hand, the appropriate choice of ν is crucial, as large cell sizes increase the false detections, while low ones enlarge the number of missing objects notably.

We can numerically compare POM to the proposed 3DMPP method in Table I, considering both test sequences and the GPE & PPE error metrics. Here in all cases the parameters have been set to minimize TER, while the corresponding FDR, MDR and MIR values are also listed. Results confirm again the superiority of the proposed 3DMPP model over POM, and also the practical equivalence of the GPE and PPE evaluation approaches. Furthermore, a second test has been performed on the two datasets, and reducing the number of cameras to two. The results are summarized in Table II.

Finally, we visualize an additional advantageous feature of the proposed 3DMPP model. In Sec. III-B, we have utilized in parallel *open* leg and *closed* leg features to correctly detect both standing and walking pedestrians. This step is a novelty over our earlier solutions [5], [6], where only the *closed* feature has been adopted. The improvement can be followed in Fig. 15, where we have backprojected the estimated ground positions on the first camera view and drawn a line between the ground plane and the estimated height. We can observe that with the *open* leg feature, the center line of the person is notably more accurate.

TABLE I
COMPARISON OF THE POM AND THE PROPOSED 3DMPP MODELS WITH OPTIMIZED PARAMETER SETS (SO THAT THE TOTAL ERROR RATE TER IS MINIMIZED), ALL THREE CAMERAS ARE USED

Sequence	Method	Ground Position Errors (GPE)			
		TER	FDR	MDR	MIR
City center	POM	**0.252**	0.179	0.073	0.000
	Prop. 3DMPP	**0.122**	0.020	0.096	0.006
Terrace	POM	**0.686**	0.354	0.331	0.001
	Prop. 3DMPP	**0.131**	0.043	0.083	0.005
Sequence	Method	Projected Position Errors (PPE)			
		TER	FDR	MDR	MIR
City center	POM	**0.205**	0.150	0.055	0.000
	Prop. 3DMPP	**0.107**	0.014	0.087	0.006
Terrace	POM	**0.607**	0.307	0.300	0.000
	Prop. 3DMPP	**0.140**	0.046	0.089	0.005

TABLE II
ROBUSTNESS ANALYSIS UNDER NON-OPTIMAL CIRCUMSTANCES: COMPARISON OF THE POM AND THE PROPOSED 3DMPP MODELS WITH USING ONLY TWO CAMERAS

Sequence	TER (GPE)		TER (PPE)	
	POM	3DMPP	POM	3DMPP
City center	0.267	0.309	0.206	0.220
Terrace	0.845	0.370	0.749	0.316

V. Conclusion

In this paper we presented a novel method to localize people in multiple calibrated cameras. For this tasks we extracted pixel-level features based on the physical properties of the 2-D image formation, and produce 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 evidence 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 [1] to extract foreground pixels. For evaluation we used the images of public semi-outdoor and outdoor datasets. According to our experiments, the proposed method produces accurate estimation, even in a cluttered environment, where full or partial occlusion is present. The output of the proposed method can be incorporated into a tracking system, which can be used to eliminate false detections.

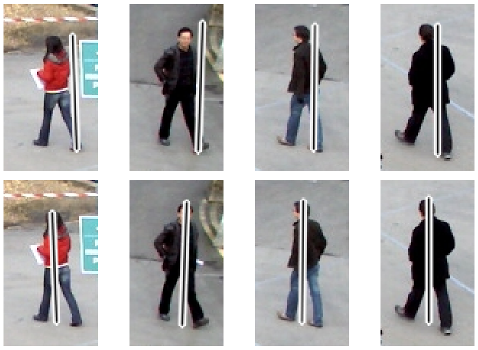


Fig. 15. Center lines of the detected cylinders projected to the images. Top: results of [5] which uses the *closed leg* ground features only. Bottom: results by using both ground features in the proposed model.

Another possible improvement might be the use of a robust body part detector (*e.g.* [28]) for creating evidence. This can be easily integrated in the proposed algorithm with minimal modification.

ACKNOWLEDGEMENTS

This work has been performed within the Transport Hub Intelligent video System (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 partially supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the Hungarian Research Fund (OTKA #101598).

REFERENCES

[1] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[2] W. Wang, J. Yang, and W. Gao, "Modeling background and segmenting moving objects from compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 670–681, 2008.

[3] C-C. Chiu, M-Y. Ku, and L-W. Liang, "A robust object segmentation system using a probability-based background extraction algorithm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 518–528, 2010.

[4] M. P. Murray, B. Drought, and R. C. Kory, "Walking patterns of normal men," *The Journal of Bone & Joint Surgery*, vol. 46, no. 2, pp. 335–360, 1964.

[5] Á. Utasi and C. Benedek, "Multi-camera people localization and height estimation using multiple birth-and-death dynamics," in *Proceedings of The 10th International Workshop on Visual Surveillance*, Queenstown, New Zealand, 2010, pp. 74–83.

[6] Á. Utasi and C. Benedek, "A 3-D marked point process model for multi-view people detection," in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3385–3392.

[7] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.

[8] I. Mikic, S. Santini, and R. Jain, "Video processing and integration from multiple cameras," in *Proceedings of the DARPA Image Understanding Workshop*, Monterrey, Canada, 1998, pp. 183–187.

[9] A. Mittal and L. S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," *International Journal of Computer Vision*, vol. 51, no. 3, pp. 189–203, 2002.

[10] J. Kang, I. Cohen, and G. Medioni, "Tracking people in crowded scenes across multiple cameras," in *Proceedings of the 6th Asian Conference on Computer Vision*, Jeju Island, Korea, 2004.

[11] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.

[12] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 505–519, 2009.

[13] D. Lingrand, "Particular forms of homography matrices," in *Proceedings of the 11th British Machine Vision Conference*, Bristol, UK, 2000, pp. 596–605.

[14] C. Benedek and T. Szirányi, "Bayesian foreground and shadow detection in uncertain frame rate surveillance videos," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 608–621, 2008.

[15] X. Descombes and J. Zerubia, "Marked point processes in image analysis," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 77–84, September 2002.

[16] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[17] X. Descombes, R. Minlos, and E. Zhizhina, "Object extraction using a stochastic birth-and-death dynamics in continuum," *Journal of Mathematical Imaging and Vision*, vol. 33, no. 3, pp. 347–359, 2009.

[18] F. Chatelain, X. Descombes, and J. Zerubia, "Parameter estimation for marked point processes. Application to object extraction from remote sensing images," in *Proceedings of The 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, Bonn, Germany, 2009, pp. 221–234.

[19] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 2913–2920.

[20] W. Ge and R. T. Collins, "Crowd detection with a multiview sampler," in *Proceedings of the 11th European Conference on Computer Vision*, Hersonissos, Heraklion, Crete, Greece, 2010, pp. 324–337.

[21] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[22] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 33–50, 2012.

[23] A. Gamal-Eldin, X. Descombes, and J. Zerubia, "Multiple birth and cut algorithm for point process optimization," in *Proceedings of The 6th International Conference on Signal-Image Technology and Internet-Based Systems*, Kuala Lumpur, Malaysia, 2010, pp. 35–42.

[24] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny, "Structural approach for building reconstruction from a single DSM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 135–147, 2010.

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

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[27] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955.

[28] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *International Journal of Computer Vision*, vol. 82, no. 2, pp. 185–204, 2009.

Ákos Utasi received the M.Sc. degree in computer sciences in 2005 from the Faculty of Information Technology, University of Pannonia, Veszprém and the Ph.D. degree in visual surveillance in 2012 from the Doctoral School of Information Science and Technology, University of Pannonia, Veszprém. He is currently a research associate with the Distributed Events Analysis Research Laboratory, at the Computer and Automation Research Institute of the Hungarian Academy of Sciences. His research interests include visual surveillance, motion detection, event detection, and action recognition.

Csaba Benedek received the M.Sc. degree in computer sciences in 2004 from the Budapest University of Technology and Economics (BME), and the Ph.D. degree in image processing in 2008 from the Pázmány Péter Catholic University, Budapest. Starting from October 2008, he worked for 12 months as a postdoctoral researcher with the Ariana Project Team at INRIA Sophia-Antipolis, France. He is currently a senior research fellow with the Distributed Events Analysis Research Laboratory, at the Computer and Automation Research Institute of the Hungarian Academy of Sciences. His research interests include Bayesian image segmentation and object extraction, change detection, 3D point cloud processing and remotely sensed data analysis.