Multi-View People Surveillance Using 3D Information

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

Ákos Utasi, Csaba Benedek, Tamás Szirányi
Computer and Automation Research Institute, Hungarian Academy of Sciences, Hungary
{utasi, bcsaba}@sztaki.hu, sziranyi@lutra.sztaki.hu

Abstract

In this paper we introduce a novel surveillance system, which uses 3D information extracted from multiple cameras to detect, track and re-identify people. The detection method is based on a 3D Marked Point Process model using two pixel-level features extracted from multi-plane projections of binary foreground masks, and uses a stochastic optimization framework to estimate the position and the height of each person. We apply a rule based Kalman-filter tracking on the detection results to find the object-to-object correspondence between consecutive time steps. Finally, a 3D body model based long-term tracking module connects broken tracks and is also used to re-identify people.

1. Introduction

The detection, localization and tracking of pedestrians are key issues in intelligent surveillance systems. The location and the trajectory of people are used in many applications, such as pedestrian counting, behavior analysis, or abnormal activity detection. However, detection becomes difficult in outdoor environment, where the monitored scenes are cluttered and the occlusion rate between the pedestrians and other static or moving objects (e.g. waving trees, traffic signs) is high. If the cameras are fixed and the object size is still limited with respect to the entire image, background subtraction is a widely used technique to separate moving objects. However, it faces two main problems in real conditions. First, since the local foreground and background color domains may partially overlap, the resulted masks of the moving objects may break apart. Secondly, due to occlusion, pixels corresponding to different objects can be merged in the same connected blobs of the motion masks. To handle the above challenges, multi-view approaches [8, 10, 19] have recently been proposed. The method in [8] uses a discretized grid on the ground plane, and assumes that the people have approximately a uniform height. [10] attempts to obtain a configuration which explains the observed data with a minimal number of occlusions, expecting that people should not be occluded in all views. Both methods [8, 10] attempt to match the complete projections of the proposed object silhouettes to the observed foreground masks, thus they strongly depend on the quality of the background subtraction step. Similarly to [19], we purely focus on the head and leg regions, when we calculate simple pixel-level features from the projections of foreground pixels on multiple parallel planes. However, we distinguish two different gait phases and derive separate descriptors to indicate pedestrians with closed and open legs, respectively. Finally, the optimal configuration of people is obtained by a stochastic birth-death process [5].

Frame-by-frame detections need to be temporally matched by means of a discriminative tracking system. To this aim we propose to use a two-stage approach, similarly to the two levels approach by Mitzel et al [15]. The first stage contains a rule based tracking system, which exploits geometrical information only (3D position and trajectory). Since occlusions and perspective problems are intrinsically solved by the detection stage, the short-term tracking performances are reliable enough [20], even if it leads to over-segmentation (i.e., the complete trajectory of some people is broken into two or more parts). However, the second stage is used to perform a long-term tracking, both connecting broken tracks and re-identifying people. Our proposal is based on the simplified 3D body model proposed by Baltieri *et al* [2], which embeds both geometrical and appearance information.

People Re-identification is a fundamental task for the analysis of long-term activities and behaviors by specific persons or to connect interrupted tracking. Algorithms have to be robust in challenging situations, like widely varying camera viewpoints and orientations, varying poses, rapid changes in clothes appearance, occlusions, and varying

lighting conditions. The first studied re-identification problem was related to traffic analysis [13] for long-term vehicle tracking, where objects are rigid, they move in well defined paths and they have mostly uniform colors. People re-identification however requires more elaborate methods in order to cope with the widely varying degrees of freedom of a person's appearance. Various algorithms have been proposed in the past: a first category of person re-identification methods rely on biometric techniques, such as face [3] or gait [12], but high resolution or PTZ cameras are required in this case. Other approaches suppose easier operative conditions, calibrated cameras and precise knowledge of the geometry of the scene: the problem is then simplified by adding spatial and/or temporal constraints and reasoning in order to greatly reduce the candidate set [14]. Finally, most re-identification methods purely rely on appearance-based features; a comparison and evaluation of some of them is reported in [7, 11]. For example, Farenzena et el [7] proposed to divide the person appearance into five parts using a rule based approach to detect head, torso and legs and image symmetries to split torso and leg regions into left and right ones. For each region, a set of color and texture features are collected and used for the matching step. Recently, Alahi *et al* [1] proposed a general framework for simultaneous tracking and re-detection by means of a grid cascade of dense region descriptors. Various descriptors have been evaluated, like SIFT, SURF and covariance matrices, and the latter are shown to outperform the formers. Finally, [9] proposed the concept of Panoramic Appearance Map to perform re-identification. This map is a compact signature of the appearance information of a person extracted from multiple cameras, and can be though of as the projection of a person appearance on the surface of a cylinder.

Our contribution is two-fold. First, we improved the localization accuracy of an existing people detection method by using an additional pixel-level feature. According to our tests, this additional step does not decrease the processing performance significantly, while it improves the accuracy by approx. 5%. Second, we applied a 3D human body model based tracking module on the frame-by-frame detections to generate the trajectories of the walking pedestrians.

2. Proposed System

The synchronized streams of input frames are processed by the people detection module, which integrates the information of all the views in order to detect people and to estimate their frame by frame position on the ground plane. A short-term tracking system is exploited to locally match the extracted detections using geometrical information and spatial constraints only. The short-term tracking parameters and thresholds should be selected to generate reliable trajectories to the detriment of their length. Finally, the long term tracking match and merge together the trajectories that

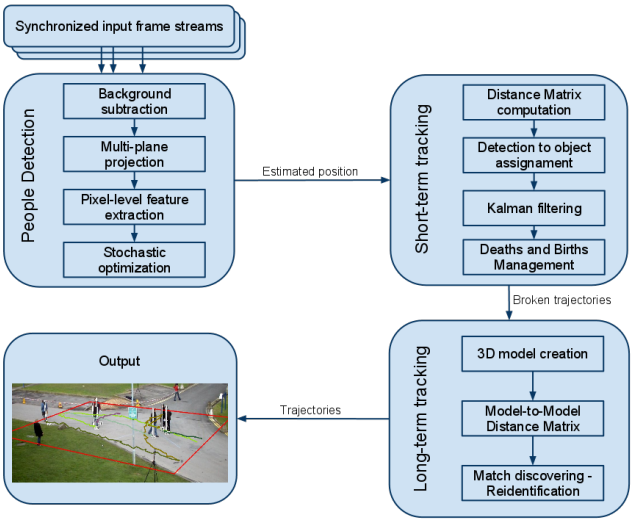


Figure 1. Work-flow of the combined tracking system

are recognized to belong to the same person. The overall system can thus be decomposed into three main modules, namely **People Detection**, **Short-term Tracking**, and **Long-term Tracking**, as depicted in Fig. 1. Section 2.1 discusses the 3D people detection method, including the feature extraction and the stochastic optimization steps. The short-term Kalman-tracker is presented in Section 2.2. Finally, the 3D body model based re-identification module is presented in Section 2.3.

2.1. People Detection

The proposed method operates in a multi camera system, and its inputs are the Tsai's calibration parameters [18] and the foreground masks extracted from each view using a Mixture of Gaussians (MoG) background model [17]. The key idea of this step is to simultaneously project the foreground pixels on the ground plane, and on a parallel plane shifted to the estimated height of the person, see Fig. 2. If this estimation is correct, we can observe from a birds-eye viewpoint that the point of osculation of the silhouette's ground and head plane projections is the ground position of the person. Since the heights of the people are unknown, we project the masks on multiple planes having distances from the ground in the range of typical human sizes. Then we fuse the projections from multiple views, and search for the optimal configuration in an iterative process using the above features and geometrical constraints.

2.1.1 Feature extraction

Our hypothesis on the location and height of a person is based on the 2D image formation of a 3D object in the conventional pinhole camera model. Let us consider in Fig. 2 the person with height h, and project the silhouette on the

P_0 ground plane (marked with blue) and on 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.

Based on the above observation we define a numerical feature, which evaluates a given $[\mathbf{p}, h]$ object candidate. We denote by $\mathbf{r}_0^i(\mathbf{p})$ a unity vector, which points from $\mathbf{p}$ towards the ground position of the ith camera on the P_0 plane, and by $\mathbf{r}_\varphi^i(\mathbf{p})$ the rotation of $\mathbf{r}_0^i(\mathbf{p})$ with angle φ. We denote the foreground points of the ith view projected to the P_0 and P_h planes by A_0^i (blue in Fig. 2) and A_h^i (red), respectively.

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 pixels on P_h (*i.e.* red prints) in the neighborhood of the $\mathbf{p}$ point in the $\mathbf{r}_0^i(\mathbf{p})$ direction, but penalize such silhouettes points in the opposite direction $\mathbf{r}_\pi^i(\mathbf{p})$. To measure this property, we define circular sectors S_h^+ and S_h^- 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, which are parameters of the model. Then, following Fig. 3(a) and (d), we calculate the *head* feature as:

$$f_h^i(p) = \frac{\mathbf{Area}\big(A_h^i \cap S_h^+(p)\big) - \mathbf{Area}\big(A_h^i \cap S_h^-(p)\big)}{\mathbf{Area}\big(S_h^+(p)\big)} \,.$$

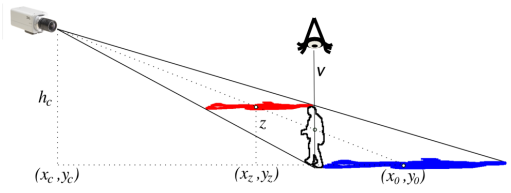
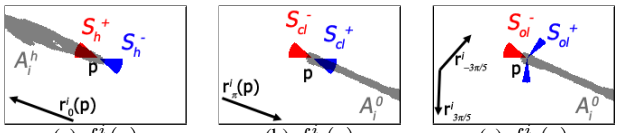


Figure 2. The available camera calibration model is used for projecting the moving body silhouettes on the ground plane (blue) and on parallel planes (red) having different heights, source: [19].



(a) $f_h^i(p)$ (b) $f_{cl}^i(p)$ (c) $f_{ol}^i(p)$

(d) $f_h^i(p)$ (e) $f_{cl}^i(p)$ (f) $f_{ol}^i(p)$

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

On the other hand, we distinguish two different cases by the definition of the *leg* position constraint. People with *c*losed *l*egs can be handled in an analogous manner to the *head* feature (see Fig. 3(b)). Here S_{cl}^+ and S_{cl}^- sectors correspond to $\mathbf{r}_\pi^i(\mathbf{p})$ and $\mathbf{r}_0^i(\mathbf{p})$ directions respectively, and

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

However, if the person is in the swing phase of the gait cycle the previous descriptor proves to be inaccurate (see Fig. 3(e)). Instead, we have developed an *o*pen *l*eg feature (see Fig. 3(c) and 3(f)), whose attractive region, S_{ol}^+, 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}^- is constructed in the same way as S_{cl}^-. Then, $f_{ol}^i(p)$ feature term is derived similarly to $f_{cl}^i(p)$. Since we have observed that for our purposes, the gait phase of each person can be fairly approximated either by the *c*losed or by the *o*pen leg states, the *joint leg* feature is obtained as $f_l^i(p) = \max(f_{cl}^i(p), f_{ol}^i(p))$. Finally, the *head* and *leg* features are truncated to take values in the $[0, \hat{f}]$ range, and are normalized by $\hat{f}$, which controls the ratio required to produce the maximal output.

If the object defined by the $[\mathbf{p}, h]$ parameters is completely visible for the ith camera, both the $f_h^i(\mathbf{p})$ 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. Therefore we construct a stronger feature by averaging the responses of the N available cameras: $\bar{f}_h(\mathbf{p}) = 1/N \cdot \sum_{i=1}^N f_h^i(\mathbf{p})$, $\bar{f}_l(\mathbf{p}) = 1/N \cdot \sum_{i=1}^N f_l^i(p)$. Finally, the joint data feature $f(\mathbf{p}, h)$ is derived as $f(\mathbf{p}, h) = \sqrt{\bar{f}_h(\mathbf{p}) \cdot \bar{f}_l(\mathbf{p})}$.

2.1.2 3D Marked Point Process model

Since the goal of the proposed model is position and height estimation of the people, we approximate a person by a cylinder u in the 3D scene, with a fixed radius R. The free parameters of the cylinder object are the center coordinate $\mathbf{p}$ on P_0 and the height h, *i.e.* $u = (\mathbf{p}, h)$ We aim to extract a configuration of n cylinder objects in the scene: $\omega = \{u_1, \ldots, u_n\}$ where n is also unknown.

We refer to the global input data with $\mathcal{D}$ in the model which consists of the foreground masks and the calibration matrices. We introduce an input-dependent energy function on the configuration space: $\Phi_\mathcal{D}(\omega)$, which assigns a *negative likelihood* value to each possible object population, and is divided into data dependent $J_\mathcal{D}$ and prior I parts:

$$\Phi_\mathcal{D}(\omega) = \sum_{u \in \omega} J_\mathcal{D}(u) + \gamma \cdot \sum_{\substack{u,v \in \omega \\ u \sim v}} I(u, v) \,, \qquad (1)$$

where $J_\mathcal{D}(u) \in [-1, 1]$, $I(u, v) \in [0, 1]$ and γ is a weighting factor between the two terms. The $u \sim v$ relation holds

if the two cylinders intersect. We derive the optimal object population as the maximum likelihood configuration estimate, *i.e.* $\omega_{\text{ML}} = \text{argmin}_{\omega\in\Omega}\left[\Phi_{\mathcal{D}}(\omega)\right]$.

In the next step, we should define the I prior and $J_{\mathcal{D}}$ data-based potential functions appropriately so that the ω_{ML} configuration efficiently describes the 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 3D model space:

$$I(u,v) = \mathbf{Area}\big(u\cap v\big)/\mathbf{Area}\big(u\cup v\big)\ . \qquad (2)$$

On the other hand, the $J_{\mathcal{D}}(u)$ *unary* potential characterizes a proposed object candidate segment u depending on the image data, but independently of other objects. Cylinders with negative unary potentials are called *attractive objects*. Based on (1) 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 in the Marked Point Process (MPP) model. Let us remember, that the f_u fitness function evaluates a person-hypothesis for u, 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 $Q(f_u, d_0)$ function: $J_{\mathcal{D}}(u) = Q(f_u, d_0) = 1 - f_u/d_0$, if $f_u < d_0$; $\exp\left(D^{-1}\cdot(f_u - d_0)\right) - 1$ otherwise. Here the d_0 parameter defines the minimal value required for acceptance. Consequently, object u is attractive according to the $J_{\mathcal{D}}(u)$ term iff $f_u > d_0$, where the d_0 parameter defines the minimal value required for acceptance.

Since finding the optimal configuration according to (1) is NP-hard, we need to use quicker optimization techniques. We have chosen the Multiple Birth and Death (MBD) algorithm [5] for this purpose, which evolves the population of people-cylinders by alternating randomized object generation (birth) and removal (death) steps in a simulated annealing framework, see details in [5, 19].

2.2. Short-term People Tracking

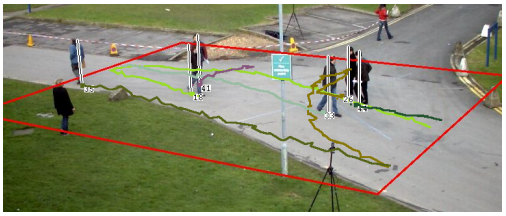
The output of the detection stage is the set of detections $\{u_n^t\}; n \in [1 \dots N^t]$ where N^t is the number of detected objects at time t. The short-term tracking system, instead, aims at creating and keeping updated a set of moving objects $\{o_p\}$. The current and future state of each object is estimated by means of a constant velocity Kalman filter. At each frame, a distance matrix between current detections and tracked objects is computed and, after a thresholding step, passed to a zero/one integer programming formulation for the assignments. The detection-to-object distance is computed using the Euclidean distance in the 3D space of the position and height of each object. The distance threshold has been set to a very low value in order to avoid wrong matches even if an over segmentation of the trajectories is introduced and handled by the long-term tracking system.

Unmatched detections are used to create new tracks only if they are localized in an entering area (to prune the wrong multiple detections which can be found in the center of the scene). Tracks without a matching detection, instead, are kept alive and updated using the Kalman prediction only. After a predefined time of inactivity or if their position exits from the scene the objects are definitively deleted. Fig.4(a) reports a qualitative example of the short-term tracking, with people id, position and trajectory superimposed. The red rectangle represents the region of interest (ROI).

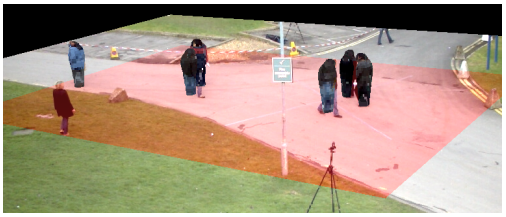
2.3. Long-term People Tracking

Broken trajectories and people entering again the scene after a while are managed by the long-term tracking algorithm. To this aim we adopted the 3D body model by Baltieri et al [2]. They proposed a monolithic 3D model, similar to a mummy's sarcophagus, which is simple enough to be processed in real time, and which embeds color appearance features useful for the re-identification stage. A new model instance Γ^p is created for each tracked person o_p obtained as in Sect. 2.2. The model $\Gamma^p = (h^p, \{v_i^p\})$ contains the person height h^p (as extracted by the detection module) and a vertex set $\{v_i^p\}$.

For the sake of completeness, let us report a brief description of the 3D body model. The model was obtained by sampling $M = 628$ vertices from a human-like surface.



(a)



(b)

Figure 4. (a) Estimated positions and heights are represented by a line. The ids and trajectories are also superimposed using different color. The red area corresponds to the ROI. (b) The 3D body models are placed in the estimated ground positions, orientation is estimated from the trajectory.

Appearance features together with some additional reliability information are stored for each vertex. In addition to the four items proposed by [2], we propose to include a saliency measure $s_i^p \in [0 \dots 1]$ for each vertex. Thus, the following features are considered (the superscript p has been neglected for the sake of clarity): (i) the normal vector $\vec{n}_i$ of the sampled surface computed at the vertex location; this feature is static and pre-computed during the manual model creation; (ii) the vertex mean color c_i; (iii) a local HSV histogram $\mathbf{H_i}$ which describes the color appearance of the vertex neighbor; it is a normalized three dimensional histogram with 8 bins for the *hue* channel and 4 bins for the *saturation* and *value* channels respectively; (iv) the optical reliability θ_i of the vertex, which takes into account how well and precisely the vertex color and histogram have been captured from the data; (v) the saliency of the vertex s_i, which indicates its uniqueness with respect to the other models; i.e., the saliency of a vertex will be high in correspondence to a distinctive logo on the person clothing and low on a common jeans patch.

2.3.1 Model creation

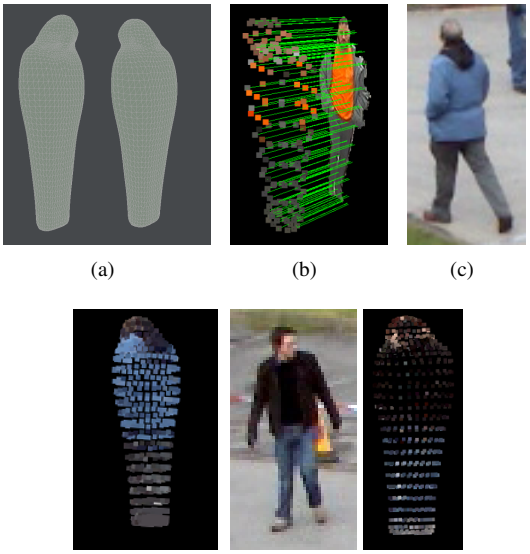
The 3D placement of the model in the real scene is obtained from the short-term tracking using the camera calibration. Assuming a vertical standing position of the people, the challenging problem to solve is the horizontal orientation of the person. To this aim, we consider that people move forward and thus we exploit the trajectory on the ground plane to give a first approximation. Given the last part of the trajectory (e.g., the last $K = 10$ positions), we try to fit a quadratic curve; if we obtain a good fit then the trajectory is classified as stable in the analyzed window and the tangent direction to the curve in the central point is assumed as the orientation of the person. A finer angle adjustment is provided by a generative approach using the already computed part of the 3D model. In Fig. 5 a sample frame and the corresponding model placement is provided: the sample positions used for the curve fitting and orientation estimation are highlighted in Fig. 5(b). An additional example is also reported in Fig. 4(b).

(a) (b)

Figure 5. 3D Model positioning and orienting: (a)the input frame; (b) estimation of the orientation from the tangent to the trajectory

Given the 3D placement and orientation of the model, each vertex is projected to the camera image plane and related to a frame pixel $x(v_i), y(v_i)$. The vertex color c_i is initialized using the image pixel upon which the vertex is projected, the histogram $\mathbf{H_i}$ is computed on a squared image patch of size N centered around $(x(v_i), y(v_i))$. The size N of the patch was selected taking into account the sampling density of the 3D model surface and the mean size of the blobs items. In our experiments, $N = 10$. Finally, the optical reliability value is initialized as: $\theta_i = \vec{n}_i \cdot \vec{p}$, where $\vec{p}$ is the normal to the image plane; the reliability gives an higher weight to front-viewed vertices and their surrounding surface rather than to lateral viewed ones. The vertices belonging to the occluded side of the person are also projected onto the image, but their reliability has a negative value due to the opposite directions of $\vec{n}_i$ and $\vec{p}$. In such a manner each vertex of the model is initialized even with a single image: from a real view if available or using a sort of symmetry-based hypothesis in absence of information. However, negative values of the reliability allow to identify vertices initialized with a forecast and not directly from the data. The vertices having no match with the current image (i.e., projected outside of the person silhouette) are iteratively initialized with a copy of the features of the nearest initialized vertex. Their reliability values however, are set to the minimum value (i.e., $\theta_i = 0$). By means of the reliability value it is possible to distinguish among vertices directly seen at least once ($\theta > 0$), vertices initialized using a mirroring hypothesis ($\theta < 0$) and vertices initialized from its neighborhood ($\theta = 0$). The described steps of the initialization phase are depicted in Fig. 6.

If multiple cameras are available or if the short-term



(a) (b) (c)

(d) (e) (f)

Figure 6. Initialization of the 3D model of a person. From left to right: the generic model, projection of the model vertices to the image, a sample frame and the corresponding 3D models from the PETS dataset

tracking system provides more detections for the same object, the 3D model could integrate all the available frames. For each of them, after the alignment step, a new feature vector is computed for each vertex successfully projected inside the silhouette of the person. The previously stored feature vector is then averaged or overwritten with the new one, depending on the signs of the reliabilities. In particular, direct measures (positive values of θ) always overwrite forecasts (negative values of θ), otherwise they are merged.

2.3.2 Object matching using 3D models

Focusing on the set $\{\Gamma^p\}$ of 3D models generated starting from the short-term tracked objects, the re-identification step aims at finding correspondences between pairs of models. First of all, a rule based selection criteria assures that candidates pairs fulfills temporal and spatial constraints (e.g., the individuals have been detected in the scene at the same time). A compatibility ranking of the remaining pairs is obtained by means of a model-to-model distance measure, which is based on the comparison of each corresponding vertex:

$$d(v_i^p, R) = d_{He}\left(\mathbf{H_i^P}, \mathbf{H_i^R}\right) = \\ = \sqrt{1 - \sum_{h,s,v} \sqrt{H_i^p(h,s,v) \cdot H_i^R(h,s,v)}} \quad (3)$$

The final score is the weighted average of the vertex-wise distances, using the product of the two reliabilities as weight:

$$D_H(\Gamma^p, \Gamma^t) = \frac{\sum_{i=1\dots M} \left(w_i \cdot d(v_i^p, v_i^t)\right)}{\sum_{i=1\dots M} (w_i)} \quad (4)$$

$$d(v_i^p, v_i^t) = d_{He}\left(\mathbf{H_i^P}, \mathbf{H_i^t}\right), \; w_i = f(\theta_i^p) \cdot f(\theta_i^t) \quad (5)$$

One of the main limitation of the proposed distances is that each vertex has the same importance and the weights w_i are based only on optical properties of the projections or the reliability of the data. Global features are useful to reduce the number of candidates (e.g., "I'm looking at people dressed with white shirt and blank pants"); the final decision, however, should be guided by original patterns and details, as humans normally do to recognize people without biometric information (e.g., a logo in a specific position of the shirt). To this aim we have enriched the vertex feature vector v_i^p with a saliency measure $s_i^p \in [0 \dots 1]$ as anticipated. Given a set of body models, the saliency of each vertex is related to its minimum distance from all the corresponding vertices belonging to the other models:

$$s_i^p \propto \min_t \left(d_H(\mathbf{H_i^P}, \mathbf{H_i^t})\right) + s_0 \quad (6)$$

where s_0 is a fixed parameter to give a minimum saliency to each vertex. The saliences s_i^p are normalized such that they sum up to 1. If s is low, the vertex appearance is similar to the one of other models and it is not distinctive; otherwise, the vertex has completely original properties and it could be used as a specific identifier of the person. A saliency-based distance can be formulated embedding the saliency in the weights of Eq. 5:

$$w_i' = f(\theta_i^p) \cdot f(\theta_i^R) \cdot s_i^p \quad (7)$$

and obtaining a corresponding saliency-based distance $D_S(\Gamma^p, \Gamma^s)$. The final distance measure used for re-identification is the product of the two distances $D_H \cdot D_S$: the first term assures the correspondence of the color distribution while the second one of the specific details. Two 3D models are classified as belonging to the same person (re-identification match) if they fulfill temporal and spatial constraints (i.e., they are not simultaneously detected in two different positions) and if their model distance is below a fixed threshold $\bar{\lambda}$.

3. Experiments

We used the publicly available *PETS* outdoor dataset [16] and the *EPFL Terrace* indoor dataset [6] to evaluate the proposed method. From the database we selected the *City center* sequence with three overlapping camera views, and manually selected a 12.2m $\times$ 14.9m ROI, which is visible from all cameras. Background subtraction was performed with the MoG [17] technique in the CIE L*u*v* color space, parameters were initialized by Expectation-Maximization [4]. During the evaluation of the proposed method the following parameters were fixed. In the feature extraction step (Sec. 2.1.1) the sector radius was set to $r = 25$cm, the angle range was constant 30°, and the feature dynamic range parameter was $\hat{f} = 0.8$. As for the parameters of the MBD optimization process, we followed the guidelines provided in [5], and used $\delta_0 = 20000$, $\beta_0 = 50$, and geometric cooling factors $1/0.96$. A sample frame from *View 001* is reported in Fig. 4(a).

For visualizing the results, we backprojected the estimated ground positions on the first camera view and draw a line between the ground plane and the estimated height (see Fig. 8). We performed visual evaluation by counting the missed detections (*MD*, the number of human bodies, that were not detected), the false detections (*FD*, the number of detections appear at positions, which are not occupied by a person), and the multiple instances (*MI*, the number of people localized multiple times in the same video frame at different positions). 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*. Finally, we calculated the total error rate: $TER = MDR + FDR + MIR$. We assumed that at least one view should correctly contain the feet and another one the head of a person, which

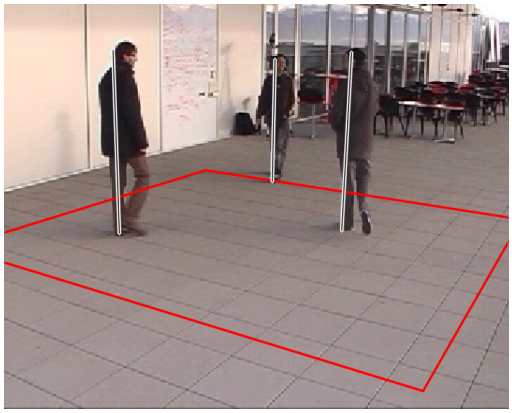


Figure 7. Example output of our people detector using the *EPFL Terrace* dataset[6].

implies a $d_0 = 1/3$ object acceptance parameter. However, due to the noisy foreground images we evaluated the $d_0 \in \{1/3.0, 1/3.25, 1/3.5\}$ set in our experiments, and we selected the d_0 parameter where the *TER* was minimal ($1/3.25$ in case of [19], and $1/3.0$ for the combined feature model). In both cases we obtained $TER \approx 10\%$. In our second experiment we evaluated the first 1000 frames of the *EPFL Terrace* dataset, and we obtained $TER \approx 7\%$ (Fig. 7 demonstrates an example output of the detector).

Next, we evaluated the localization accuracy of the proposed method using the combination of f_{cl} and f_{ol}, and compared it to the [19] model which is purely based on the f_{cl} feature. We carefully counted the number of successful detections where the localization accuracy of the two methods was significantly different (see Fig. 8 for examples). According to our experiments in 80.40% of the cases the combined feature model produced better results. Expressing in percent of the number of all people detections counted in the 400 frames we obtained 5.04% improvement over [19].

The short-term tracking system parameters were selected in order to minimize errors at the expense of oversegmentation of tracks. For this reason, the Kalman filter based tracker does not introduce particular errors and the provided detections are correctly handled and linked, without any id-exchange or missing detection. Thus, the corresponding numerical results are not reported in this section. However, since the same people enter and exit the rectangular ROI and since almost all tracks have been oversegmented by the short term tracker, the long-term tracking system is used to detect correspondences among tracks. During tracking, a 3D model was created for each track. Not all frames were used to initialize and update the model's appearance features: only those with the highest overlap between the 3D model backprojection and the foreground were automatically chosen. Then saliency measures were computed between all the model created so far and re-identification was performed. The long term tracking system was able to correctly link most of the tracks, obtaining precision and recall values of 72.73% and 88.8% respectively.

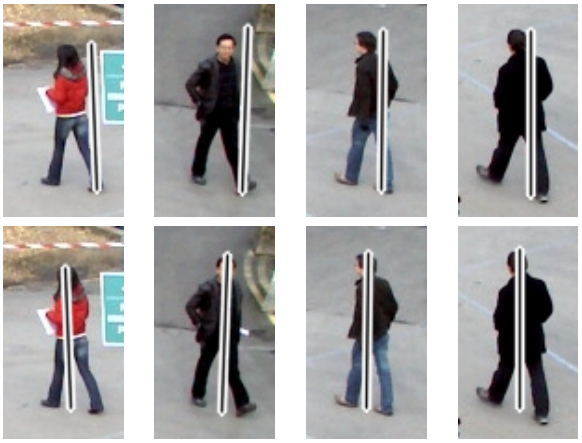


Figure 8. Center lines of the detected cylinders projected to the images. Top: results of [19] which uses the *closed* ground features only. Botton: results by using both ground features in the proposed model.

4. Conclusions

In this paper we presented a novel system for visual surveillance applications. The main novelty of our approach is that we use 3D information to detect, track and re-identify pedestrians. Moreover, we improved the localization accuracy of a state-of-art method by using an additional feature. The proposed method has been tested on a public dataset, and according to our experiments it achieves accurate results in cluttered outdoor environment.

5. Acknowledgements

This work is currently under development within the project THIS (JLS/2009/CIPS/AG/C1-028), with the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security.

References

[1] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, 2010. 2

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. 3D body model construction and matching for real time people re-identification. In *Proc. of Eurographics Italian Chapter Conference (EG-IT 2010)*, Genova, Italy, Nov. 2010. 1, 4, 5

[3] M. Bäuml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen. Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In *Proc. of AVSS*, 2010. 2

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Society, Series B*, 39(1):1–38, 1977. 6

[5] 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. 1, 4, 6

[6] EPFL. Dataset - Terrace Sequence, 2008. http://cvlab.epfl.ch/data/pom/. 6, 7

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of CVPR*, pages 2360–2367, June 2010. 2

[8] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. on PAMI.*, 30(2):267–282, 2008. 1

[9] T. Gandhi and M. Trivedi. Panoramic Appearance Map (PAM) for Multi-camera Based Person Re-identification. In *Proc. of AVSS*, pages 78–78, Nov. 2006. 2

[10] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *Proc. of the 11th European Conf. on Computer Vision*, 2010. 1

[11] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *Proc. of PETS 2007*, 2007. 2

[12] L. Havasi, Z. Szlavik, and T. Sziranyi. Eigenwalks: walk detection and biometrics from symmetry patterns. In *Proc. of ICIP*, pages III–289, 2005. 2

[13] T. Huang and S. Russell. Object identification: A bayesian analysis with application to traffic surveillance. *Artificial Intelligence*, 103:1–17, 1998. 2

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

[15] D. Mitzel, E. Horbert, A. Ess, and B. Leibe. Multi-person tracking with sparse detection and continuous segmentation. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (1)*, volume 6311 of *Lecture Notes in Computer Science*, pages 397–410. Springer, 2010. 1

[16] PETS. Dataset - Performance Evaluation of Tracking and Surveillance, 2009. http://www.cvg.rdg.ac.uk/PETS2009/. 6

[17] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22(8):747–757, 2000. 2, 6

[18] 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. 2

[19] A. Utasi and C. Benedek. Multi-camera people localization and height estimation using multiple birth-and-death dynamics. In *Workshop on Visual Surveillance*, 2010. 1, 3, 4, 7

[20] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13, 2006. 1