Scale Invariant 3D Multi-Person Tracking using a Base Set of Bundle Adjusted Visual Landmarks

Alberto Del Bimbo Giuseppe Lisanti Federico Pernici
{delbimbo, lisanti, pernici}@dsi.unifi.it

University of Florence – MICC – Italy

Abstract

We present a vision system for real-time 3D tracking of multiple people moving over an extended area, as seen from a rotating and zooming camera. Despite the general problems of multiple target tracking (MTT), the use of a pan–tilt–zoom (PTZ) camera adds several difficulties for the multiplicity of connected problems.

Our approach exploits multi-view image matching techniques to index and refine, at runtime, the closest world to image homography for the current view. This is made possible by applying (in a batch phase) bundle adjustment method over a set of distinctive visual landmarks extracted from the field of regard [1] of the zooming camera sensor.

The approach is experimentally evaluated on several difficult video sequences. Quantitative results show that the proposed approach makes it possible to deliver stable tracking performance in scenes of previously infeasible complexity. We achieve an almost constant standard deviation error of less than 0.3 meters in recovering 3D trajectories of multiple moving targets in an area of 70x15 meters.

1. Introduction

In recent years computer vision has seen tremendous progress and several algorithms have become applicable for real-world tasks. These successes have promoted demand for active vision systems than can autonomously operate in scenarios of daily life. This is interesting for applications such as wide area scene monitoring using robotic cameras. The goal of wide area monitoring, gaze redirection and zoom control on target details, has opened new problems for tracking, especially for use as surveillance devices. In particular, abnormal behavior detection at a distance demands both 3D trajectories analysis [20] and the necessary image resolution to perform biometric recognition [21]. This cannot generally be achieved with a single stationary camera sensor. Even though a large number of stationary or PTZ cameras were adopted, making them operate in a cooperative way is an expensive solution. Making the most use of a single zooming sensor is a worthy goal.

In this paper we focus on tracking multiple people over an extended area as observed by a single rotating and zooming camera sensor. Such a scenario puts notable demands on two well known classes of vision algorithms: multi view geometry for camera sensor registration (i.e. manage visual data to account for camera zooming and pose change) [10] and multiple target tracking [28].

In order to recover 3D metric trajectories of multiple targets at a distance, camera parameters must be taken into account during the measurement process. It is well known that any metric measurement in the scene needs some form of camera calibration. For zooming cameras, however, precalibration is almost impossible, since it is difficult to recreate the full range of zoom and focus settings. Markers are not available, and 3D measurements are not easily obtained for points visible in the provided scene. Known uncalibrated techniques are currently far from being real time and the task of aligning the systems coordinate frame, with that of the tracked objects remains non-trivial. However, a correct exploitation of zooming lens could provide high accuracy for tracking targets at a distance. In particular, the accuracy at which a target position can be estimated depends on the distance between the target and the camera (the more a target is distant from the camera, the larger the measurement uncertainty is). High zooming-in induces high accuracy in the measurement. To the best of our knowledge no studies have taken advantage of this potential accuracy to recover multiple trajectories in wide areas.

In addition, tracking of multiple targets with a zooming camera sensor becomes even more challenging because of the imaged scale variations of the targets due to change of camera focal length, camera redirection and change of object to camera distance. False measurements extraction due to wrong scale inference could prevent also data association since the computational complexity increases exponentially

[1] The camera field of regard is defined as the union of all field of view over the entire range of pan and tilt rotation angles and zoom values.

with clutter [26]. This because we don't know if a given observation was a false alarm or a correct measurement from a target.

1.1. Relationship to Previous Work

In the literature multiple target tracking with a moving camera follows two distinct approaches: 1) performing target localization from an adequate representation of the target shape and appearance 2) performing target localization from target dynamics and registration of the moving sensor.

Several partial solutions have been proposed. Some authors have considered the problem of managing the target scale change especially when tracked with zooming cameras. In [19], the authors combine the boosting detector of [27] with tracking based on particle filters. No camera motion or geometric scene information are used in this case; imaged target scale is estimated according to the boosting detector (the target scale dynamics is modeled in its state and estimated through a particle filter). The boosting learning process requires off-line acquisition of target appearances.

In [29], this work is extended by taking into account also the sensor movements. In particular, the world-to-image homography computed from the hockey rink model is considered [18]. In this way, a more complex dynamic model is specified in 3D world coordinates. Scale dynamics is not modeled in the target state and the mean-shift algorithm [3] is used to stabilize the results of tracking to predict better the location of the target. Adaptation to scale changes is performed by examining windows slightly larger/smaller than the current target size.

In [24], the authors perform tracking by modeling the whole image as a set of layers. Each layer consists of an elliptic shape, a target motion model (translation, rotation and scale) and layer appearance (intensity modeled using a single Gaussian). Sensor registration is performed by compensating the background motion using the estimated inter-image homographies. In this way, the target motion can be estimated from the compensated image. The approach has been applied to airborne vehicle tracking and to activity monitoring, with ground-based stationary pan-tilt-zoom camera. In the latter case target size changes more than in the airborne tracking; larger shape variance is needed to accommodate size changes.

Recent advances in the image searching techniques have been shown that real-time landmark matching and localization with a large landmark database has become possible [17]. Based on this technique, the work [9] have been shown that a query in a 1 million image database takes 0.02 seconds. This results in a frame-rate of about 50Hz, well suited for real-time robot global localization. Despite this success, no focus is given to take advantage of this effort aimed to multiple target tracking with robotic zooming cameras. The closest work related to our approach is [7] in which discriminative features from a single view are used to perform image object transfer in a collaborative camera network.

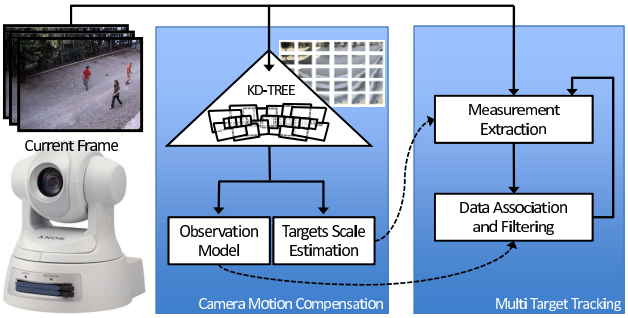

Figure 1. *Left*: Off the shelf PTZ camera. *Right*: Components of our system and their connections, executed for each frame of a video sequence.

1.2. Overview and paper contribution

Our proposed approach differs from previous similar work on MTT using a single PTZ camera [19, 29, 24] mainly in three aspects: object scale inference, observation model estimation and EKF-CJPDAF filtering model.

We exploit multi-view image matching to recover and refine at runtime the closest world to image homography and the closest focal length with respect to the current view. Estimation is carried out by indexing a set of bundle adjusted [25] visual landmarks extracted from the field of regard of the zooming camera sensor. Under the assumption of vertical stick like targets moving on planar scene, the *target scale* and the *observation model* are directly estimated from the target state and the geometric relationships between the PTZ camera and the single view geometry of viewed scene. This permits to obtain very effective and accurate template matching for target measurement extraction. Indeed scale knowledge significantly reduces the expected computational cost of template matching and increases accuracy in measurement extraction; while the observation model computed by linearizing the time-variant world to image homography correctly captures the measurement uncertainties as the camera moves and zooms. Based on this evidence, we adopt an EKF-CJPDAF tracking and data association filter that carefully combines the interplay between the two modules.

We provide several new contributions in this research: (1) We introduce a new multiple target tracking framework for a single rotating and zooming camera that combines the robustness of tracking-by-detection, the accuracy of a batch bundle adjustment optimization and the efficiency of EKF-CJPDAF with scale invariant target template matching; (2) Differently from any previous work performing MTT with a single PTZ camera, we are able to recover 3D metric trajectories of moving persons with almost constant uncertainty

less than 0.3 meters, at more than 70 meters distance from the camera; (3) Differently from [29], no explicit 3D model based alignment is performed to obtain the time variant world to image homography since we match unstructured features. Moreover our batch learning strategy is simpler since it does not require learning target appearance.

Fig. 1 gives an overview of the proposed vision system. Each frame is processed as follows: first, a nearest neighbor image is retrieved from the bundle adjusted image set and a new camera pose is estimated; then targets measurements are extracted taking advantage of motion prediction and scale information inferred from camera pose and zoom value. At any point in time the system uses only information from the last and current frame. The rest of the paper is organized as follows. Sections 2 and 3 detail the three main components of the system, namely observation model estimation, scale invariant target measurement extraction and multiple target tracking. Finally sect. 4 presents experimental results.

2. Estimating the Observation Model

In the case of a PTZ camera viewing a plane ($z = 0$), the observation model must assume a non-linear function $\mathbf{g}_t$ relating at time t the targets world coordinates $(x_t, y_t, 0) \in \mathbb{R}^3$ to image measurement $\mathbf{z}_t \in \mathbb{R}^2$:

$$\mathbf{z}_t = \mathbf{g}_t(x_t, y_t) + \mathbf{v}_t \tag{1}$$

where $\mathbf{v}_t$ is an uncorrelated stochastic process with Gaussian distribution $\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{V})$. $\mathbf{V}$ is the 2×2 covariance matrix of the measurement noise which models image pixellation noise and image localization errors, and holds constant. Since we have assumed that targets move in a plane, $\mathbf{g}_t$ is the time variant world-to-image homography that models the projection from the world plane to the current camera view.

Estimating the time variant world to image homography $\mathbf{g}_t$ with a PTZ camera in real-time is a challenging problem. Recent works identify features in the images for which the real-world coordinates can be measured (i.e. they perform model based alignment) [11, 18]. The main advantage of these methods is the absence of drift because the system is always registering directly to the scene as long as a subset of the features is visible. This strategy has also the benefit that the coordinate system is always implicitly defined. However, in many real-world cases, especially with zooming cameras, three critical inherent drawbacks may occur: 1) the 3D reference model is difficult to obtain, 2) the subset of the features is not always visible because of zoom in operations, 3) measurable 3D features may not be easily detectable.

A general solution for these problems is to compute an initial homography between a ground plane in a frame and the world ground plane [22]. Then, detect the interest

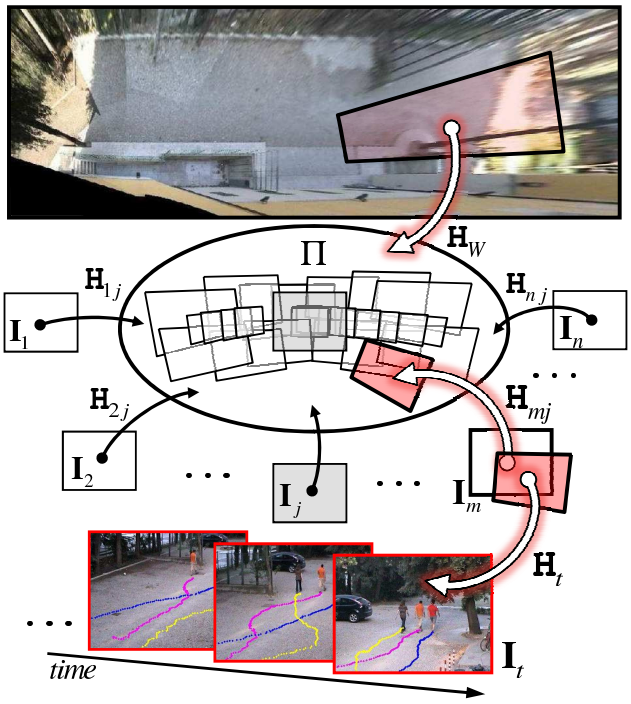

Figure 2. Camera motion detection and camera geometry indexing. The current image $\mathbf{I}_t$ is matched and its closest camera geometry retrieved through $\mathtt{H}_{mj}$ (see text).

points in the reference frame and the next frame, check their matching, and hence use the resulting correspondences to compute the next homography. The process is repeated for all subsequent frames. However the clear limitation of such approach is that concatenating transformations introduces accumulation of small errors. Frame to frame registration leads to misalignments, which invalidate the observation model of eq. 1, and results in a bias in target measurement. Some recent works limit this drifting effect by performing real-time global optimization over several frames [2, 13]. A main drawback is that all these methods are limited to the special case of panning and tilting smooth motions. No explicit general zooming camera motions are considered.

We avoid these drawbacks by matching unstructured features that index bundle adjusted scene geometry. This *tracking by detection* approach considerably facilitates the model based alignment problem, avoiding complex recursive filtering approach or global optimization at run-time [14]. Obviously, this can be done if it is possible to train the system beforehand. According to this, differently from [7] which uses just one image, during a learning stage a base set of visual landmark points from some reference views is build. The reference views are taken so as to cover the whole field of regard of the PTZ system at several different pan tilt angles and zoom settings. Each image in the base set has an ensemble of descriptors that correspond to location features seen from different camera poses and focal length values.

Each view $\mathbf{I}_{1..n}$ is supplemented with a bundle adjusted

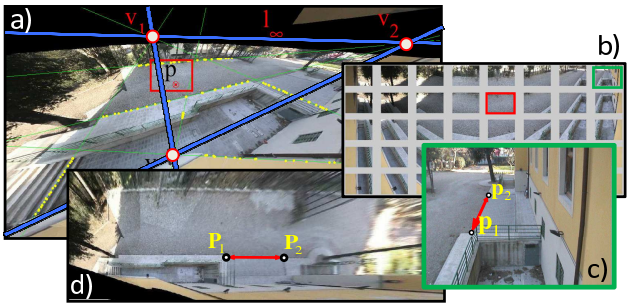

Figure 3. a): Three orthogonal vanishing points in the reference plane. b): The base set images. The red rectangle indicates the reference image c): Imaged points with 3D known length. d): The rectified reference plane.

homography, $\mathtt{H}_{lj}$ with $l = 1..n$, that relates the view in the base set to a common reference plane Π (i.e. the planar mosaic plane) as shown in fig. 2. The reference plane Π is related to 3D world coordinates through homography $\mathtt{H}_W$. Each homography $\mathtt{H}_{lj}$ is parameterized by two rotation angles (pan and tilt) and the focal length. Optimization is carried out by minimizing the global reprojection error according to bundle-adjustment [6].

At runtime the match for a SIFT feature extracted from the current frame $\mathbf{I}_t$ is searched according to the Euclidean distance of the descriptor vectors. Because the keypoints are detected in scale–space, the scene does not necessarily have to be well–textured which is often the case of planar man–made scene. SIFT based matching exploits scale invariance and is therefore appropriate in the presence of camera zooming operations, moreover SIFT is also indicated for blurred features due to its multi-resolution character.

To allow for fast search the base set is organized as a KD–tree. The search is performed so that bins are explored in the order of their closest distance from the query description vector, and stopped after a given number of data points have been considered [16]. In particular since the number of images that may overlap in a single ray is small (we assume $n_o = 4$) each feature is matched to its n_o nearest neighbors. These nearest neighbor features may belong to different images of the base set and a vote for the corresponding view is taken if the descriptors distance ratios satisfy: $\frac{d_{k\text{–NN}}}{d_{(k+1)\text{–NN}}} < 0.67,\ k = 1 \ldots n_o - 1$. The image $\mathbf{I}_m$ closest to the current view $\mathbf{I}_t$ is the one having the greatest number of feature matches (i.e. votes). Once $\mathbf{I}_m$ is found, the homography $\mathtt{H}_t$ relating $\mathbf{I}_t$ to $\mathbf{I}_m$ is computed at run time with RANSAC (see fig. 2). The clear advantage is that wrong feature matches, being distributed over a large number of images, are less likely to cause an incorrect image match. This allows for a significant reduction in the number of RANSAC iterations required at runtime.

Finally the world to reference plane transformation $\mathtt{H}_W$ is obtained exploiting single view geometry properties of the base set (see fig. 3(a) and fig. 3(b)). The mosaic of the base set is obtained from the inter-image homographies $\mathtt{H}_{lj}$ between a reference image $\mathbf{I}_j$ and each of the other images in the base set. From this single wide view, the transformation $\mathtt{H}_W$ maps any point in the 3D world plane onto the reference plane Π. The 3D scene plane is mapped onto the current image $\mathbf{I}_t$ as (see Fig. 2):

$$\mathtt{G}_t = \mathtt{H}_t \mathtt{H}_{mj}^{-1} \mathtt{H}_W \tag{2}$$

The world to mosaic homography $\mathtt{H}_W$ is computed as $\mathtt{H}_W = \mathtt{H}_p^{-1}\mathtt{H}_s$ where $\mathtt{H}_p$ is the rectifying homography defined as: $\mathtt{H}_p = \begin{pmatrix} \beta^{-1} & -\alpha\beta^{-1} & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & 1 \end{pmatrix}$, the scalars l_1, l_2, α and β are obtained from the projections of the circular points of the imaged scene plane [15]. $\mathtt{H}_s$ is a similarity transformation obtained from the 3D length of two specified imaged points $\mathbf{p}_1$, $\mathbf{p}_2$ as shown in fig. 3(c) and fig. 3(d).

2.1. Target scale inference

The target imaged scale can be directly estimated from the target state and from the geometric relationships between the PTZ camera views and the matched image onto the base set. This allows to obtain a robust and more effective estimation of the target image likelihood.

Because targets have been assumed to be closely vertical in the 3D scene plane, they can be approximated by a rectangular bounding box template in the image. The position of the two extremities (the imaged feet and heads location for humans) are related by a planar homology [4]. This transformation is obtained by exploiting information about the directions orthogonal and parallel to the scene plane, namely the vanishing point and the vanishing line of the plane. Since for the individual images of the sequence, the vanishing line and the vanishing points change according to the variation of the camera parameters due to the pan-tilt-zoom operation, for each image at time t the planar homology constraint is expressed as:

$$\mathtt{W}_t = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_{t,\infty} \cdot \mathbf{l}_{t,\infty}^{\mathrm{T}}}{\mathbf{v}_{t,\infty}^{\mathrm{T}} \cdot \mathbf{l}_{t,\infty}}\ , \tag{3}$$

and is completely defined by $\mathbf{l}_{t,\infty}$ and $\mathbf{v}_{t,\infty}$ obtained respectively as:

$$\mathbf{l}_{t,\infty} = \mathtt{G}_t \cdot [0,\ 0,\ 1]^{\mathrm{T}}, \qquad \mathbf{v}_{t,\infty} = \mathtt{K}_t \mathtt{K}_t^{\mathrm{T}} \cdot \mathbf{l}_{t,\infty} \tag{4}$$

The cross-ratio μ, being projective invariant, remains constant throughout the sequence. The internal camera matrix $\mathtt{K}_t$ is computed in closed form by directly exploiting the homography $\mathtt{H}_t$ in eq. 2. This is achieved by solving for $\mathtt{K}_t$ the following equation:

$$\mathtt{H}_t\, \mathtt{K}_m\, \mathtt{K}_m^{\mathrm{T}}\, \mathtt{H}_t = \mathtt{K}_t \mathtt{K}_t^{\mathrm{T}} \tag{5}$$

The internal camera parameters $\mathtt{K}_m = \begin{pmatrix} f_m & 0 & x_0 \\ 0 & f_m & y_0 \\ 0 & 0 & 1 \end{pmatrix}$ are retrieved from the closest image $\mathbf{I}_m$ as described in the

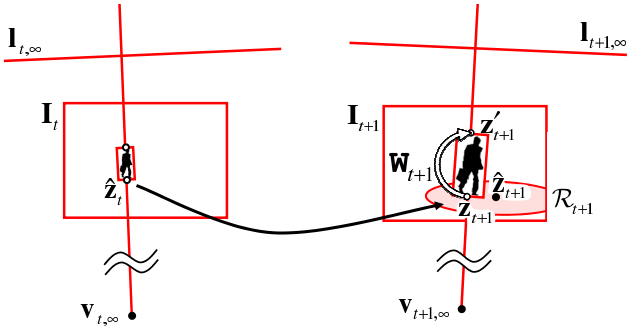

Figure 4. Measurement extraction process in the neighborhood of the predicted measurement $\hat{\mathbf{z}}_{t+1}$. The measurement $\mathbf{z}_{t+1}$ (foot position) is derived by transforming the template according to $\mathtt{W}_{t+1}$. $\mathbf{z}'_{t+1}$ is the head position and $\mathcal{R}_t$ is the measurement region.

previous section. Assuming $\mathtt{H}_t = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}$ and $\mathtt{K}_t = \begin{pmatrix} f_t & 0 & x_0 \\ 0 & f_t & y_0 \\ 0 & 0 & 1 \end{pmatrix}$ we have two equations for the focal length f_t:

$$f_t^2 = \frac{f_m^2\,(h_{11}^2 + h_{12}^2) + h_{13}^2}{f_m^2\,(h_{31}^2 + h_{32}^2) + h_{33}^2}, \quad f_t^2 = \frac{f_m^2\,(h_{21}^2 + h_{22}^2) + h_{23}^2}{f_m^2\,(h_{31}^2 + h_{32}^2) + h_{33}^2} \tag{6}$$

We take the mean of the two constraints of eq. 6. This estimation has a very good stability and accuracy since the focal length f_m in the intrinsic camera matrix $\mathtt{K}_m$ for the matched image $\mathbf{I}_m$ is estimated offline, as described in sect. 2, and takes into account the whole base set because of the bundle adjustment optimization.

The planar homology constraint can be applied at runtime using the eq. 3 to estimate the predicted imaged scale of the target template at time $t+1$. For each target, measurement search is computed as follows: we assume $\mathbf{z}_{t+1}$ and $\mathbf{z}'_{t+1}$ to be respectively the imaged extremities of a stick-like target, specified in the image $\mathbf{I}_{t+1}$; these two points are related through $\mathtt{W}_{t+1}$ of eq. 3 according to:

$$\mathbf{z}'_{t+1} = \mathtt{W}_{t+1}\mathbf{z}_{t+1}. \tag{7}$$

If the predicted measurement $\hat{\mathbf{z}}_{t+1}$ is considered, the predicted head position is estimated applying eq. 7 to the target template in $\hat{\mathbf{z}}_{t+1}$. This operation is performed in the neighborhood of $\hat{\mathbf{z}}_{t+1}$. In this case the target measurement position is obtained by searching for the maximum of the image likelihood (see fig. 4). Unlike the iterative approaches presented in previous works our scale inference formulation is analytic and indirectly takes into account the geometry and the appearance of the whole field of regard.

3. Multiple person tracking

The scale inference strategy just described in the previous section allow us to perform scale invariant template matching inside the target search region. This is achieved by adopting color spatiograms template matching since they retain information about the geometry of object feature distributions [1] (color histogram alone would not benefit of any prior knowledge about target scale).

According to this we use a color spatiogram template $h_{\mathcal{T}}(b) = \langle n_b, \mu_b, \Sigma_b \rangle$, $b = 1, ..., B$ with $B = 4$ taken in the first frame of the sequence. The values of n_b, μ_b and Σ_b are respectively the number of pixels in the b-th bin, the mean and covariance of the coordinates of those pixels.

According to this, differently from [24, 19], it is possible to decouple the target imaged size from its position and speed so that a simplified target state model can be used. Therefore, the following model is assumed:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_{t-1}, \tag{8}$$

where the target state is modeled as $\mathbf{x} \in \mathbb{R}^4$, $\mathbf{x}_k = [x_t, y_t, \dot{x}_t, \dot{x}_t]$, being $\dot{x}_t, \dot{y}_t$ the target velocity components, $\mathbf{A}$ the constant velocity model transition matrix; $\mathbf{w}_t$ is an uncorrelated stochastic process with Gaussian distribution $\mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q}_t)$ and $\mathbf{Q}_t$ is the 4×4 covariance matrix of the process noise, that accounts for the deviations from the assumed behavior (i.e. target maneuvers).

The process noise $\mathbf{w}_t = \mathcal{N}(0, \mathbf{Q})$, and particularly covariance $\mathbf{Q}$, must therefore model the capability of the target to change acceleration and/or motion direction between consecutive instants. If we assume $\mathbf{w}_t$ to be constant during the k-th sampling period, the increase in the velocity during this period is $w_t \Delta t$, while the effect of this acceleration on the position is $w_t \Delta t^2 \frac{1}{2}$. The following expression for covariance $\mathbf{Q}$ is therefore obtained as:

$$\mathbf{Q} = \begin{pmatrix} \sigma_x^2 \frac{1}{4}\Delta t^4 & 0 & \sigma_x^2 \frac{1}{2}\Delta t^3 & 0 \\ 0 & \sigma_y^2 \frac{1}{4}\Delta t^4 & 0 & \sigma_y^2 \frac{1}{2}\Delta t^3 \\ \sigma_x^2 \frac{1}{2}\Delta t^3 & 0 & \sigma_x^2 \Delta t^2 & 0 \\ 0 & \sigma_y^2 \frac{1}{2}\Delta t^3 & 0 & \sigma_y^2 \Delta t^2 \end{pmatrix}. \tag{9}$$

To have a realistic representation of the target movement, in that targets move mainly in the lateral directions, the components of target acceleration σ_x, σ_y should be different. A rotation matrix can be applied to covariance $\mathbf{Q}$:

$$\mathbf{D}_t = \begin{pmatrix} \mathbf{I}_{2\times 2} & 0 \\ 0 & \mathbf{R}_{2\times 2} \end{pmatrix}, \qquad \mathbf{R}_{2\times 2} = \begin{pmatrix} \frac{v_t}{\|v_t\|} \\ \frac{a_t}{\|a_t\|} \end{pmatrix}, \tag{10}$$

where $v_t \in \mathbb{R}^2$ is the velocity vector with component $v_t = [\dot{x}_t \;\; \dot{y}_t]^T$ and $a_t \in \mathbb{R}^2$ is the vector orthogonal to v_t, with components $a_t = [-\dot{y}_t/\dot{x}_t \;\; 1]^T$. Our formulation allows to perform the fast recursive estimation of the Extended Kalman Filter (EKF), where the only equation that is explicitly affected by the process noise is the covariance time update equation: $\mathbf{P}_t^- = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^{\mathrm{T}} + \mathbf{D}_t\mathbf{Q}\mathbf{D}_t^{\mathrm{T}}$.

The measurement equation of eq. 1 is finally linearized around the prediction $\hat{\mathbf{x}}_t^-$, according to EKF formulation as:

$$\widehat{\mathbf{G}}_t = \begin{pmatrix} \frac{\partial \mathtt{G}_t^1}{\partial x_t} & \frac{\partial \mathtt{G}_t^1}{\partial y_t} & 0 & 0 \\ \frac{\partial \mathtt{G}_t^2}{\partial x_t} & \frac{\partial \mathtt{G}_t^2}{\partial y_t} & 0 & 0 \end{pmatrix}_{\mathbf{x}=\hat{\mathbf{x}}_t^-}. \tag{11}$$

The closed form recursive estimation of the EKF is further exploited to enhance real time performance by using

an *ad-hoc* JPDAF formulation, for data association [8]. If the total number of observations and tracks is large the use of JPDAF is computationally expensive. CJPDAF (Cheap JPDAF) calculates the probability of track k being associated with measurement i as:

$$\beta_{ki} = \frac{p_{ki}}{S_k + S_i - p_{ki} + C}, \qquad (12)$$

where $p_{ki} = \mathcal{N}(\nu_i(k))$, being $\nu_i(k)$ the innovation of the k-th track wrt i-th measurement, $S_k = \sum_{i=1}^{M} p_{ki}$, $S_i = \sum_{k=1}^{T} p_{ki}$ and C is a parameter that models clutter density, T is the number of targets and M is the number of observations at time k. This technique heavily weights measurements in only one covariance target region, and lightly weights measurements that lie in an area with several overlapped covariance target regions. If several other tracks can be associated with measurement i, a large S_i will lower the weight. If the track has several measurements to choose from, all weights will be lowered by S_k. This calculation gives higher weights to those measurements which are closest to the predicted position and which are associated by the fewest number of other tracks.

4. Experimental results

Performance evaluation in 3D multiple target tracking with a zooming camera is a very complicated task. Some errors can be due to a mistake of the camera tracker rather than a weakness of the data association strategy. A person could have been undetected because of wrong scale inference, due to optical and mechanical misalignments in the lens system. It is not easy to separate the performance of the components from that of the overall system.

To address these difficulties we used a real data-set, acquired with an off-the shelf SONY SNC-RZ30 PTZ camera, deployed over an area of approximately 900 squared meters, long 70 meters. The video sequences are recorded at 368×272 pixels and 20fps. The scene is learned, as described in sect. 2, from 150 images acquired at different levels of pan, tilt and zoom, so as to cover the whole field of regard. We considered several scenarios of increasing difficulty with three and four targets. Each sequence is acquired with continuous pan, tilt and zoom (difficult sequences introduce also a zoom out step during targets occlusion). Initial target image templates are acquired manually, though automatic initialization can be achieved by plugging the detector [5] or [12] into our framework. We quantitatively measure performance by comparing generated and manually annotated trajectories. In total about 3500 frames are examined. Our system with non-optimized code (apart from the SIFT library) reaches about 20 frames per second.

Fig. 5 refers to the results obtained in the three targets scenario. Fig. 5(a) shows the rectified reference plane, with superimposed recovered trajectories and some sample

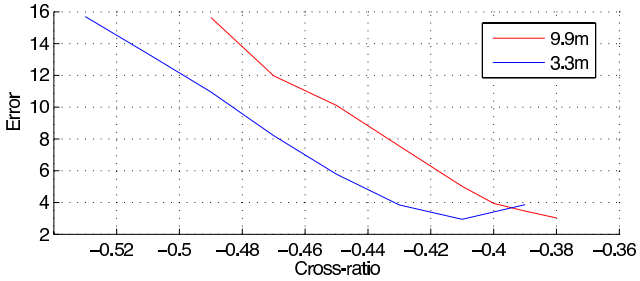

Figure 6. An example of target scale error with varying the cross-ratio μ. The plots are reported for two different measured 3D world lengths.

frames. In particular, left and bottom frames show two target occlusions resolved at different zoom factor. The top four frames indicate the adaptation of the motion model to a target maneuver. Target uncertainty increases in the direction of lateral acceleration. Fig. 5(b) shows the recovered trajectories superimposed in the reference plane. The first three plots in fig. 5(c) show the estimation error for each target in the scene. The error is small and remains constant over time as the target to camera distance and zoom factor increase. This result shows that the proposed method correctly exploits the zoom to get tracking accuracy at a distance. The fourth plot in fig. 5(c) shows the targets speed expressed in m/s; the correct estimation of speed allows to detect the two running targets. The fifth plot in fig. 5(c) shows focal length advancement from about 450 pixels to about 2000 pixels, corresponding approximately to a $4\times$ zoom factor. Fig. 5(d) confirms the correct exploitation of zoom lenses. The uncertainty (3σ error ellipses) of each target remains almost constant as targets walk away from the camera. The covariance ellipse increases only when the target is occluded because of the data association mechanism.

As one would expect, error localization is more pronounced along depth direction. Indeed the standard deviation error is higher along the camera z direction. In our experiments the average standard deviation is 0.3 meters, also when the target to camera distance is more than 70 meters. This good performance is possible because of the absence of drift, due to the continuous detection and accuracy of the associated homography ($\mathtt{H}_{mj}$ in eq. 2).

To confirm these results all sequences were quantitatively analyzed by varying two different system parameters: 1) the 3D world known length used to compute similarity transformation $\mathtt{H}_W$ in eq. 2; 2) the cross-ratio μ of eq. 3. Fig. 6 shows the target imaged scale error for two different known length in the 3D world. The tracker yields moderate error for a reasonable deviation of the cross-ratio, showing low sensitivity to this parameter. In the last experiment we test the performance of the tracker for a very challenging sequence with four targets.

Fig. 8 shows the first part of the sequence where four targets (two dressed similarly) are walking close together

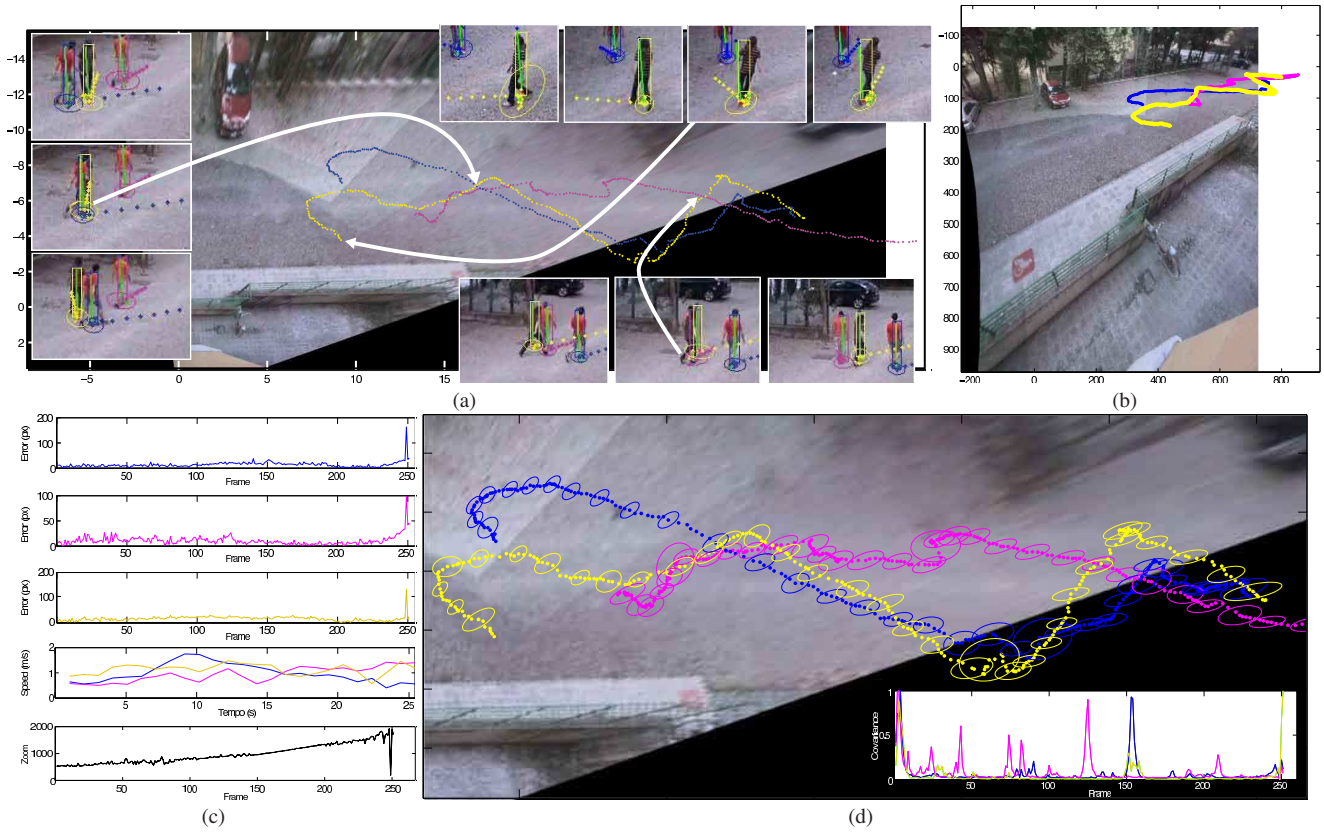

Figure 5. *(a):* The 3D trajectories recovered by our system, here superimposed onto the rectified mosaic. *(b):* The same trajectories plotted in a low resolution mosaic image. *(c):* A series of plot showing: The estimation error for the three targets, their speeds and the estimated camera focal length in pixels. *(d):* A detail of fig.(a), the recovered trajectories are here plotted with the filtered uncertainties of the 3D target location (3σ regions computed from the covariance matrix). The plot bottom-left shows the determinant of the covariance matrix.

Figure 8. Some frames extracted from a challenging tracking problem. In particular it is emphasized the intersection of all four targets and the tracker ability to handle complex scenarios. 3D Trajectories are overlayed at run-time in the video frames showing accuracy in camera parameters estimation. Besides the absence of camera parameters smoothing, overlayed trajectories does not present excessive jittering.

crossing each other; one of the targets is maneuvering, trying to steal identities to the others.

Fig. 7(*top*) reports the uncertainty of each target. When all the targets are occluded (between frame 150 and 200) the camera performs a zoom out motion. This causes a tracking failure due to low resolution of targets appearance.

5. Conclusion

In this paper we have presented a real-time system that produces long and stable tracks in complex scenarios. We improve the effectiveness of multiple target tracking systems to scene of previously infeasible complexity. Our system combines the robustness of tracking-by-detection with

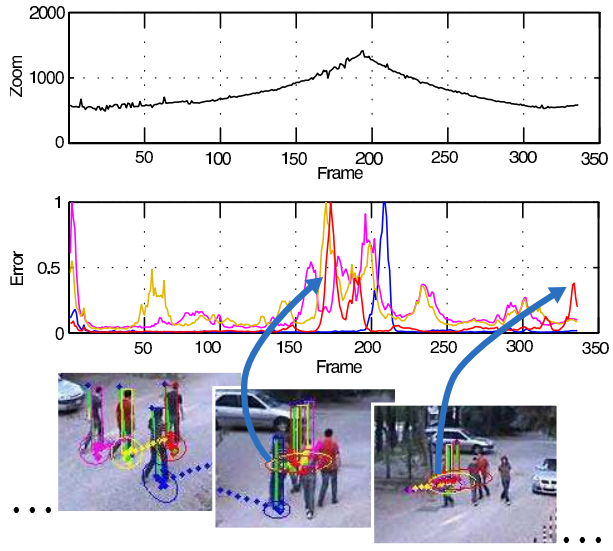

Figure 7. *Top:* focal-length advancement estimation during tracking failure. *Bottom:* The camera zooms out when targets occlude each other (between frame 150 and 200). The graph shows the covariance determinant value for each target.

the accuracy of a batch bundle adjustment optimization. Both these characteristics are exploited at runtime to infer the observation model and target scale of a single PTZ camera sensor.

Various directions are available for future research: 1) control the camera to track several targets simultaneously, slewing the video sensor from target to target and zooming in and out as necessary; 2) landmarks maintenance in continuous changing background; 3) 3D "photo-tracking" with hand-held cameras using bundle adjusted data sets as used for site-exploration in "photo-tourism" [23].

Acknowledgment. This work is partially supported by the EU IST VidiVideo Project (Contract FP6-045547) and by Thales Italia, Florence, Italy.

References

[1] S. T. Birchfield and S. Rangarajan. Spatiograms versus histograms for region–based tracking. 2005.

[2] J. Civera, A. J. Davison, J. A. Magallon, and J. M. M. Montiel. Drift-free real-time sequential mosaicing. *International Journal of Computer Vision*, 2008.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, 2003.

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, November 2000.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, 2005.

[6] L. de Agapito, E. Hayman, and I. D. Reid. Self-calibration of rotating and zooming cameras. *International Journal of Computer Vision*, 45(2), November 2001.

[7] A. Del Bimbo, F. Dini, A. Grifoni, and F. Pernici. Uncalibrated framework for on-line camera cooperation to acquire human head imagery in wide areas. In *Proc. of IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2008.

[8] R. J. Fitzgerald. Pack biases and coalescence with probabilistic data association. *IEEE Transactions on Aerospace and Electronic Systems*, AES-21, 1985.

[9] F. Fraundorfer, C. Engels, and D. Nistér. Topological mapping, localization and navigation using image collections. In *IROS*, pages 3872–3877, 2007.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[11] R. Hess and A. Fern. Improved video registration using non-distinctive local image features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[13] D.-W. Kim and K.-S. Hong. Real-time mosaic using sequential graph. *Journal of Electronic Imaging*, 15(2):023005, 2006.

[14] V. Lepetit and P. Fua. *Monocular Model-based 3d Tracking of Rigid Objects (Foundations and Trends in Computer Graphics and Vision(R))*. Now Publishers Inc, 2005.

[15] D. Liebowitz and A. Zisserman. Metric rectification for perspective images of planes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, 2004.

[17] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, June 2006.

[18] K. Okuma, J. J. Little, and D. G. Lowe. Automatic rectification of long image sequences. *The Asian Conference on Computer Vision*, 2004.

[19] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proc. European Conference on Computer Vision*, 2004.

[20] A. Prati, S. Calderara, and R. Cucchiara. Using circular statistics for trajectory analysis. In *Proc. of International Conference on Computer Vision and Pattern Recognition*, 2008.

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proc. of the 17th International Conference on Pattern Recognition*, 2004.

[22] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *Proc. International Symposium on Augmented Reality*, pages 120–128, Oct. 2000.

[23] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[24] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2002.

[25] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proc. of the International Workshop on Vision Algorithms*, 2000.

[26] J. Vermaak, S. Godsill, and P. Perez. Monte carlo filtering for multi target tracking and data association. *IEEE Transactions on Aerospace and Electronic Systems*, 41(1):309–332, Jan. 2005.

[27] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

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

[29] N. d. F. Yizheng Cai and J. Little. Robust visual tracking for multiple targets. *European Conference on Computer Vision*, 2006.