Scale Invariant 3D Multiple Person Tracking with PTZ camera

***Giuseppe Lisanti*, Alberto Del Bimbo, Federico Pernici**

Media Integration and Communication Center (MICC), University of Florence, Italy [a]

[a]This work is partially supported by Thales, Florence (Italy), http://www.thalesgroup.com/

1. Introduction

Goal: **Tracking multiple people over an extended area as observed by a single rotating and zooming camera sensor.**

Issues:

- **target scale variations** due to change of camera focal length, camera redirection and change of object to camera distance;
- camera parameters must be taken into account during the measurement process in order to recover **3D metric trajectories** and scale of multiple targets at a distance;
- **curse of dimensionality** of Multi Target Tracking that can result in a high computationally expensive task if the total number of observations and tracks is large.

Such a scenario puts notable demands on two well known classes of vision algorithms: multi view geometry for **camera sensor registration** and **multiple target tracking**.

2. The System

We propose a vision system for **real-time 3D tracking of multiple people** moving over an extended area with a **single PTZ sensor**.

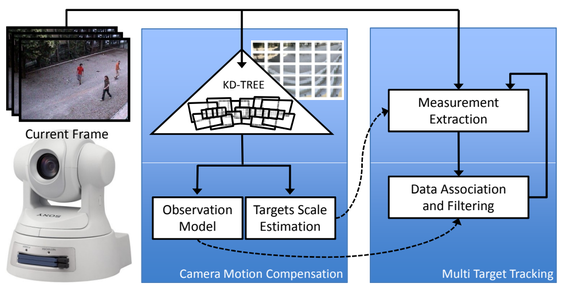

Figure 1: *Overview of our system.*

- 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.
- This is carried out by indexing a set of **bundle adjusted visual landmarks** extracted from the field of regard of the zooming camera sensor.
- 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.
- The estimated scale permits to obtain very effective and accurate template matching for target **measurement extraction**.

3. 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 $\mathbf{z}_t \in \mathbb{R}^2$: $\mathbf{z}_t = \mathbf{g}_t(x_t, y_t) + \mathbf{v}_t$.

- Each view $\mathbf{I}_{1..n}$ is supplemented with a **bundle adjusted** homography, $\mathtt{H}_{lj}$ with $l = 1..n$, that relates the view in the base set to a common reference plane Π.
- The reference plane Π is related to **3D world coordinates** through homography $\mathtt{H}_W$.
- The image $\mathbf{I}_m$ closest to the current view $\mathbf{I}_t$ is the one having the **greatest number of feature matches**. 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.
- The **3D scene** plane is mapped onto the current image $\mathbf{I}_t$ as:

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

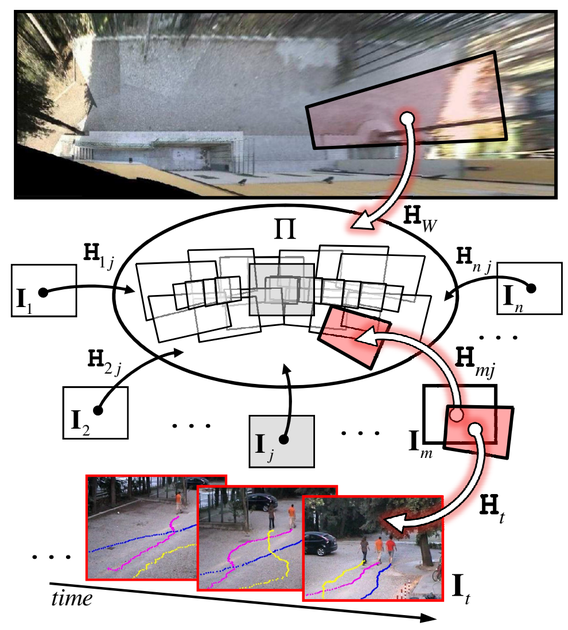

Figure 2: *Camera motion detection and camera geometry indexing.*

SIFT based matching exploits **scale invariance** and is therefore appropriate in the presence of camera zooming operations.

4. Target scale inference

Each target can be approximated by a rectangular bounding box template in the image. The position of head and feet are related by a **planar homology** $\mathtt{W}_t$. This **time-variant homology** change according to the variation of the camera parameters due to the pan-tilt-zoom operation.

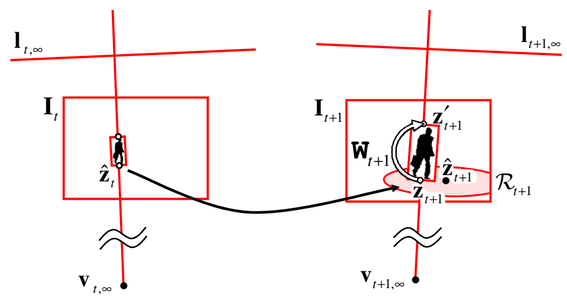

Figure 3: *Measurement extraction process in the neighborhood of the predicted measurement.*

The internal camera matrix $\mathtt{K}_t$ is computed in closed form by directly exploiting the homography $\mathtt{H}_t$ in order to obtain the **time-variant homology**.

$$\mathtt{H}_t \, \mathtt{K}_m \, \mathtt{K}_m^{\mathsf{T}} \, \mathtt{H}_t = \mathtt{K}_t \mathtt{K}_t^{\mathsf{T}} \tag{2}$$

$$f_t^2 = \frac{f_m^2 \, (h_{11}^2 + h_{12}^2) + h_{13}^2}{f_m^2 \, (h_{31}^2 + h_{32}^2) + h_{33}^2} \tag{3}$$

This estimation has a very good stability and accuracy since the focal length f_m is estimated offline with **bundle adjustment optimization**.

5. Multiple person tracking

- The scale inference strategy allows us to perform **scale invariant template matching** inside target search region.
- This is achieved by adopting **color spatiograms template matching** since they retain information about the geometry of object feature spatial distributions.
- For tracking we perform the fast recursive estimation of the **Extended Kalman Filter (EKF)**:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_{t-1}, \tag{4}$$

$$\mathbf{P}_t^- = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^{\mathsf{T}} + \mathbf{D}_t\mathbf{Q}\mathbf{D}_t^{\mathsf{T}} \tag{5}$$

- This is further exploited to enhance real time performance by using an *ad-hoc* JPDAF formulation, for data association.

  CJPDAF (Cheap JPDAF) calculates the probability of track k being associated with measurement i as:

$$\beta_{ki} = \frac{p_{ki}}{S_k + S_i - p_{ki} + C}\,, \tag{6}$$

  where $p_{ki} = \mathcal{N}(\nu_i(k))$, being $\nu_i(k)$ the innovation of the k-th track wrt i-th measurement, $S_k = \sum_{i=1}^{M} p_{ki}$, $S_i = \sum_{k=1}^{T} p_{ki}$.

- 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.

6. Results

Fig. 4 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.
The plot in fig. 4 shows **focal length advancement** from about 450 pixels to about 2000 pixels, corresponding approximately to a $4\times$ zoom factor.

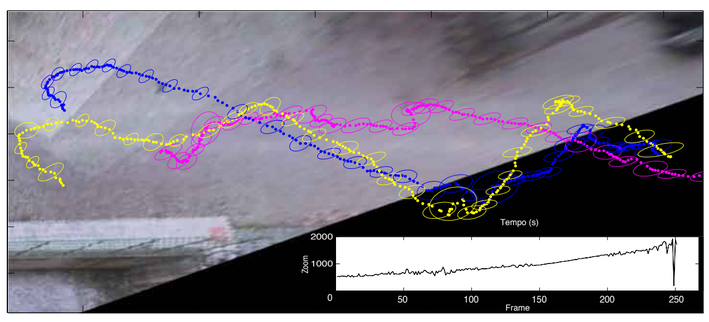

Figure 4: *The recovered trajectories are here plotted with the filtered uncertainties of the 3D target location. We are able to recover 3D metric trajectories of moving persons with **almost constant uncertainty** less than **0.3 meters**, at more than **70 meters distance** from the camera.*

Fig. 5 shows the **targets speed** expressed in m/s; here is possible to see that the correct estimation of speed allows to detect the two **running targets**.

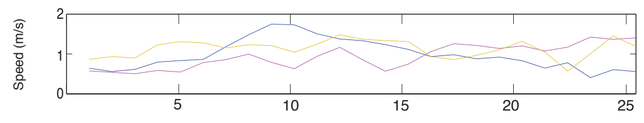

Figure 5: *The speed of three targets and the estimated camera focal length in pixels.*

Fig. 6 shows the first part of a sequence where four targets (two dressed similarly) are walking close together crossing each other; one of the targets is maneuvering, **trying to steal identities** to the others.

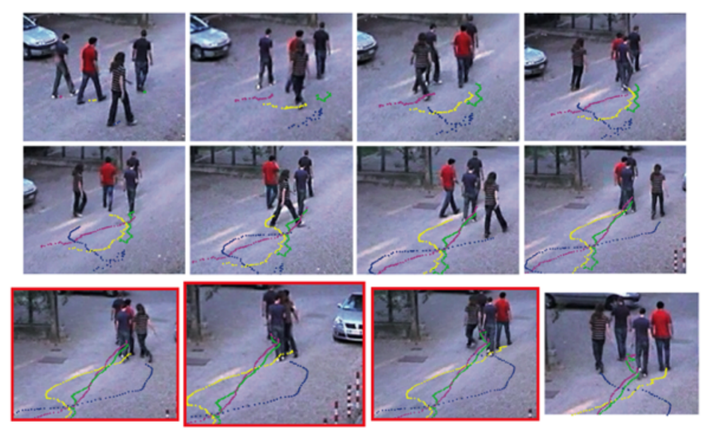

Figure 6: *Some frames extracted from a challenging tracking problem.*