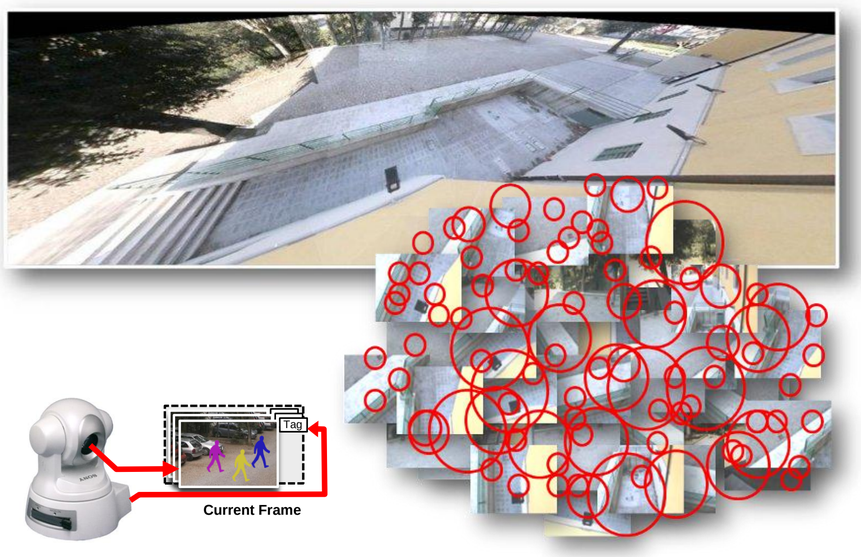
LIFELONG LOCALIZATION AND MAPPING WITH A ROTATING AND ZOOMING CAMERA

Del Bimbo A., Lisanti G., **Masi I.**, Pernici F.

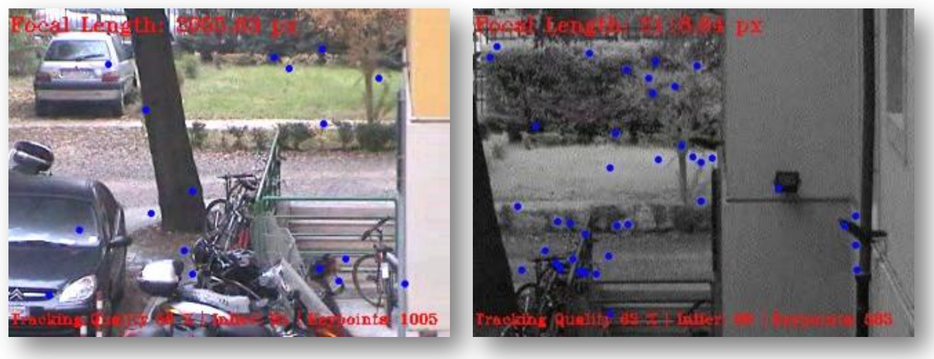
Multimedia Integration and Communication Center, University of Florence, Italy

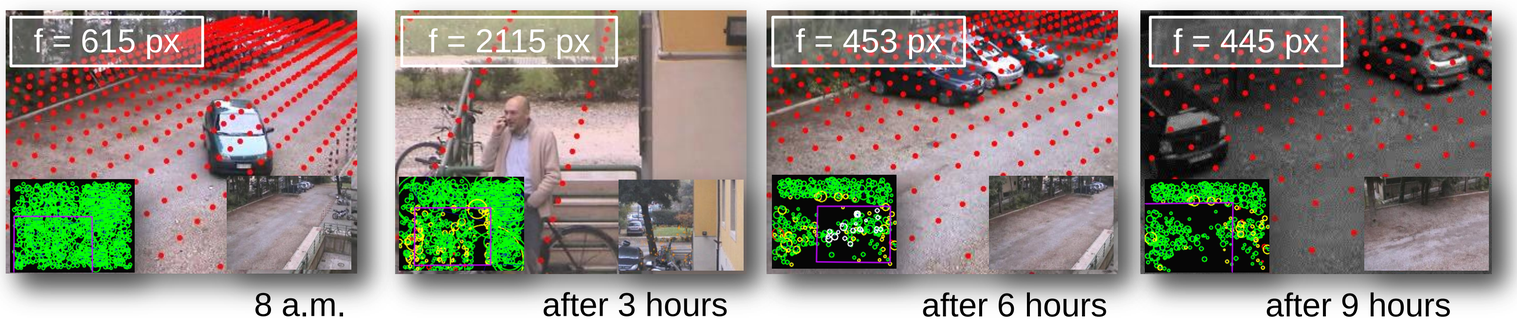
{*delbimbo,lisanti,masi,pernici*}*@dsi.unifi.it - http://www.micc.unifi.it*



Initialization: Camera geometry and scene appearance are initialized from keyframe images. Bundle adjustment optimization and SURF keypoint are used.
Textual tagged images: Keyframes and the current view are associated with a tag provided by the device.

Goal: Update camera parameters, scene geometry and appearance while preserving the bundle adjustment accuracy of the initialization.
Main Issue: Lifelong real time pose estimation with large focal length.
Results: 9 hours of continuous running time with focal lengths up to **2000px** (320x240@20FPS).





Abstract This work presents a method of estimating the pose of a single PTZ camera in a dynamic environment. While this has previously been attempted by adapting SLAM algorithms developed for robotic exploration, no explicit varying focal length estimation has been introduced before. We propose a novel system designed to track a PTZ camera in a wide area by exploiting device-tagged text information. The system indexes and refines at runtime a set of poses from a pre-build map of the observed scene.

Localization and Mapping Formulation:

The problem is that of inferring: $p(\mathbf{x}_t, \mathbf{m}_t | \mathbf{z}_{0:t}, \mathbf{u}_{0:t}, \mathbf{x}_0, \mathbf{m}_0)$ where $\mathbf{x}_t$ is the state vector describing the internal and external camera parameters; $\mathbf{m}_t$ is a variable width vector describing the locations of scene landmarks and $\mathbf{z}_t$ is a set of landmarks observed from the camera. In order to avoid recursive statistical filtering no frame to frame motion coherence is exploited, tracking failure to unmodeled motion is avoided by performing tracking by detection (i.e. visual appearance is used to recognize place similarity). According to this the formulation simplifies to: $p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_0)p(\mathbf{m}_t|\mathbf{z}_t, \mathbf{m}_0)$.
The estimation of the equation above is initialized with N keyframes each of which has attached the localization $\mathbf{x}_0 = \{\mathbf{x}_1^0, \mathbf{x}_2^0, \cdots, \mathbf{x}_\mathrm{N}^0\}$ and the initial map structure $\mathbf{m}_0$ both estimated using batch bundle adjustment optimization. For each keyframes, the system stores attached to each localization of $\mathbf{x}_0$ a tag $\mathbf{p}_i \in \mathbb{R}^3$ observed when the device is respectively localized in the state $\mathbf{x}_0$ as: $\mathbf{p} = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_\mathrm{N}\}$. At runtime a similar tag is observed from the current view $\mathbf{p}_t$ (i.e. the tag of the current PTZ actuators) and compared with those in $\mathbf{p}$ as:

$$i^\star = \arg\min_{\mathbf{p}_i \in \mathbf{p}} \|\mathbf{p}_t - \mathbf{p}_i\|_2. \quad (1)$$

Map Appearance Model:

We use an adaptive bag of features appearance model to find features in the map that are likely to have similar appearance to the current video frame. Each visual landmark descriptor is chosen in image locations given by an interest point detector. Map landmarks are also associated with a lifetime probability: $\mathbf{M}_t = \{\mathbf{m}_t, \mathbf{d}_t, p(\tau_t)\}, \quad \forall i = 1..\mathrm{N}$.
Model appearance is updated by a running average recursive filter as: $\mathbf{d}_t = (1-\alpha)\cdot\mathbf{d}_{t-1} + \alpha\cdot\mathbf{d}_t$ for each matched landmark descriptor and insertion and deletion of landmarks is performed by generating samples based on lifetime probability according to the MCMC strategy.

Real Time Localization and Mapping:

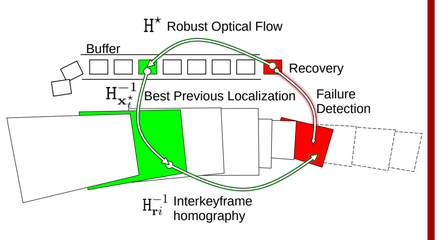
Localization is computed with respect to the reference keyframe as: $\mathrm{H}_t = \mathrm{H}_{\mathbf{x}_t^\star}^{-1}\mathrm{H}_{\mathbf{r}i}^{-1}$ where interkeyframe homographies are defined with respect to a reference keyframe $\mathbf{r}$ as $\mathrm{H}_{\mathbf{r}i}\ \forall i = 1, \cdots, \mathrm{N}$ as computed in the initialization. Under the assumption we've made the map conditional density $p(\mathbf{m}_t|\mathbf{x}_t, \mathbf{z}_t, \mathbf{m}_0)$ is computed by fusing the observations over time. Recursive estimation for a landmark location is computed as: $\mu_t = (1-t^{-1})\cdot\mu_{t-1} + t^{-1}\cdot\hat{\mathbf{m}}_t$, $\Sigma_t = (1-t^{-1})\cdot\Sigma_{t-1} + t^{-1}\cdot(\hat{\mathbf{m}}_t - \mu_t)(\hat{\mathbf{m}}_t - \mu_t)^\mathrm{T}$ where $\hat{\mathbf{m}}_t = \mathrm{H}_{\mathbf{x}_t^\star}^{-1}\cdot\mathbf{z}_t$. To promote local stationarity of each 3D back-projected rays we used an infinity time window average. This assumes a 2D Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ the landmark locations.
The measurements are assumed independent of each other and may originate from true physical landmark or from clutter (i.e. moving objects in the scene). The observation model is described in the form: $p(\mathbf{z}_t|\mathbf{m}_t) \Longleftrightarrow \mathbf{z}_t = \mathbf{h}_t(\mathbf{m}_t) + \mathbf{v}_t$ where $\mathbf{h}_t(\cdot)$ is the time varying function defining the measurements process and $\mathbf{v}_t$ is a zero mean noise process accounting for interest point localization error *and* homography estimation error. The measurement equation is evaluated by searching over the device-tagged reference views (eq. 1) and refined with RANSAC.

Failure Detection and Recovery:

Camera pose failures are detected observing the statistical distribution of inliers. Once a failure is detected, recovery procedure is started to restore the current camera pose by searching for the best previous localization obtained in past frames. Relocalization is then performed as (see figure):

$$\mathrm{H}_t = \mathrm{H}_{\mathbf{r}i}^{-1}\cdot\mathrm{H}_{\mathbf{x}_t^\star}^{-1}\cdot\mathrm{H}^\star$$

where $\mathrm{H}^\star$ is obtained by tracking features points with Lucas-Kanade algorithm.



Experimental Validation and Applications

Fig. a) The history of lifetime for landmarks of the map collected in 10 minutes of continuous operation. *Fig. b)* The two curves illustrate the number of inliers overtime. The red curve shows the performance of our proposed method while the blue curve shows the performance in absence of a keypoint maintenance strategy such as in [1]. *Fig. c)* Landmark localization error and standard deviation of keypoint $id = 3298$ (not present in the initialization). Errors denote that the accuracy of the initialization is preserved.

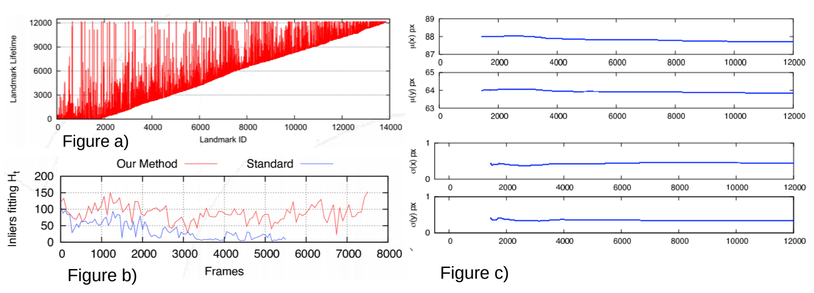


Fig. d-top) 3D grid superimposed on to the planar mosaic of the learned area with some samples from the multitarget tracking experiment [1] with our framework. *Fig. d-bottom)* The trajectories are superimposed on to the rectification of the scene mosaic.

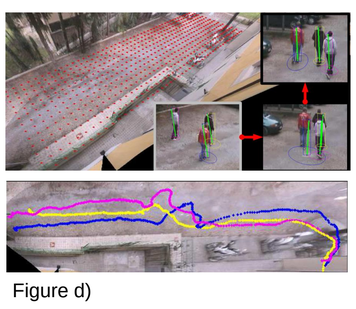


Figure d)

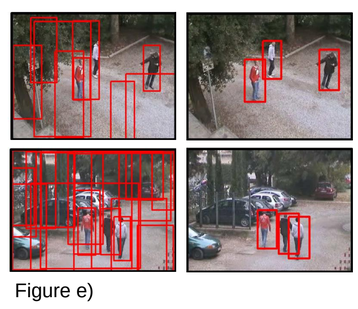


Figure e)

Fig. e-left) Part-based people detector *Fig. e-right)* Pedestrian detection results exploiting geometric context as provided by our framework [2].

References

[1] A. Del Bimbo, G. Lisanti, F. Pernici, *"Scale Invariant 3D Multi-Person Tracking using a Base Set of Bundle Adjusted Visual Landmarks"* in Proc. of ICCV Int'l Workshop on Visual Surveillance (VS), Kyoto, Japan, 2009.

[2] A. Del Bimbo, G. Lisanti, I. Masi, F. Pernici, *"Person Detection using Temporal and Geometric Context with a Pan Tilt Zoom Camera"* in Int'l Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.