

# Real-Time Tracking Avoids the Correspondence Problem in Spatio-Temporal Analysis

Christoph Stock, Axel Pinz  
Graz, University of Technology,  
Institute of Electrical Measurement and Measurement Signal Processing  
{stock, axel.pinz}@tugraz.at

## 1. Problem Statement

There are many well-known applications of spatio-temporal analysis, which require *known* landmarks (fiducials, control points, “natural” landmarks) in the scene, e.g., camera pose estimation [Lu], object tracking [Kato], and wide baseline stereo [Baumberg]. Typically we face a high initial complexity of the model-to-image or model-to-scene correspondence problem [Brandner]. Imagine for example the initialisation of a tracking process based on the matching of a simple scene model (e.g. 5 control points) and a feature set (e.g. several hundred points extracted from the first image of the sequence). This problem may be alleviated by extracting additional descriptive features (colour, cornerness, texture, etc., see [Stock]), but still persists. Recent progress in real-time tracking and online structure and motion analysis has led us to a new approach, which is based on frame-to-frame point correspondences and on emerging structure estimation, and completely avoids the hard matching problem during initialisation or re-initialisation. This opens up a new field of online spatio-temporal reasoning which may be successfully applied in several high-potential applications.

## 2. Ingredients for Practical Spatio-Temporal Analysis

**Hybrid tracking:** We have developed a new hybrid tracking system, which determines in real-time the six degrees of freedom (6 DoF) of a sensor head motion (3 translations and 3 rotations) [Ribo]. The system consists of a camera and an inertial sensor with three accelerometers and three gyroscopes. In principle, both sensor concepts could be used standalone to determine 6DoF pose information. Camera pose can be calculated from known landmarks in the scene (solving the perspective  $n$  point – PnP-problem), and inertial pose requires a calibration for gravity  $g$ , double integration for position, and single integration for orientation information. Each individual sensor has its particular breakdown characteristic: vision-based tracking fails for high motion blur (fast rotations), or when landmarks are temporary occluded. Inertial tracking fails for slow motion (poor signal to noise ratio), and requires regular offset correction to compensate sensor drift.

**Robust camera pose:** Camera pose calculation is tricky. Real-time operation demands that only a limited number of landmarks can be tracked, which limits accuracy and stability of the pose algorithm. There are well-known critical configurations, where PnP algorithms may fail. Robustness is gained by a combination of intelligent target selection, motion prediction, fast image acquisition, and precise landmark localisation. We use CMOS camera technology to access individual regions of the sensor array, which leads to update rates for individual corners of up to 2kHz, and we localize corners at subpixel accuracy.

**Online structure estimation:** If we try to find (and subsequently track) known landmarks in the scene, we face the complexity problem of correspondence search mentioned above. Having a reliable hybrid tracking system, we can initialise the first frame with a number of interest points (e.g. corners), track them and use the inertial tracker for initial motion estimation. By online estimation of the scene structure, we get a coarse and sparse (3D landmarks resembling the interest points) model of the scene. Initialisation can be further improved by using a calibrated stereo rig. Figure 1 shows our sensor head with 2 CMOS imaging sensors, an inertial tracker, and our custom hardware for direct pixel access.

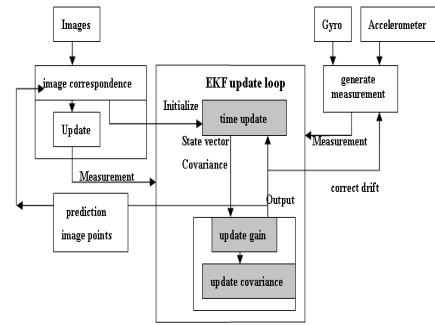
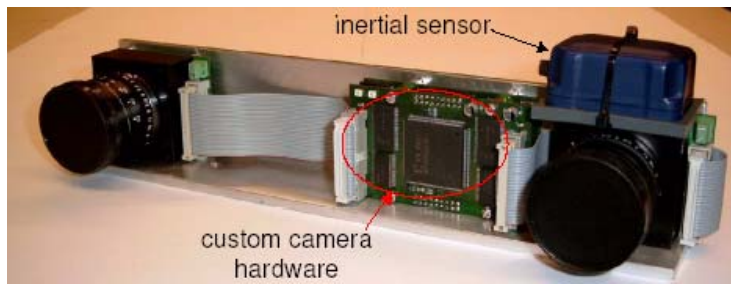


Figure 1: (a) Sensor head for real-time tracking and structure estimation, (b) Framework for sensor fusion

### 3. Sample Applications

#### 3.1 Structure and Motion by Sensor Fusion

We developed a novel framework [Chen] for real-time sensor fusion. Based on an extended Kalman filter, which consists of two independent measurement channels for vision- and inertial measurements, we are able to simultaneously determine the camera pose and to estimate the structure of the scene (see Figure 1b).

#### 3.2 Object-centred Feature Selection

Based on spatio-temporal mapping of local image features, combined with reliable camera pose information, we are able to cluster image features which are either related with objects or to the background. This yields a significant reduction of input data for further object categorization modules.

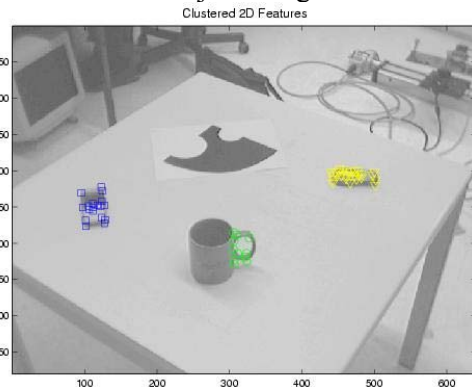
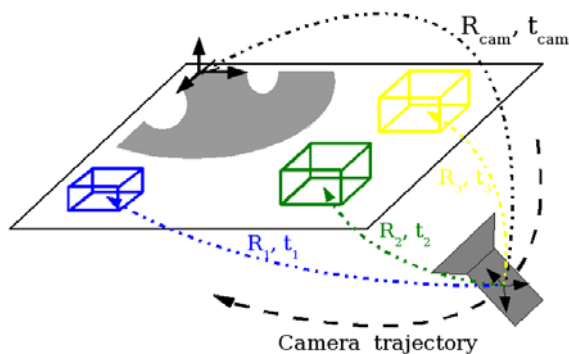


Figure 2: (a) System sketch consisting of an artificial target, several objects and a moving camera; (b) Real image with labeled object categories.

### 4. Conclusions

We show that spatio-temporal image processing, which is based on reliable tracking technologies combined with robust pose estimation can be done in real-time. This opens up a completely new field of high-level image processing applications. Additionally, we give you a brief overview on our recent work.

### References

Luet al: Fast and Globally Convergent Pose Estimation from Video Images, pp 610-622, IEEE Transactions on Pattern Analysis and Machine Intelligence 2000, Vol 22  
 Kato, H., Billinghurst, M., Popyrev, I., Imamoto, K., Tachibana, K. *Virtual Object Manipulation on a Table-Top AR Environment*. Proc. International Symposium on Augmented Reality, pp.111-119, (ISAR 2000), 2000.  
 Baumberg, A. *Reliable Feature Matching Across Widely Separated Views*. Proc. CVPR. 2000  
 Brandner M., Pinz, A. *Real-time tracking of complex objects using dynamic interpretation tree*. Proc. 24<sup>th</sup> DAGM Symposium, Springer LNCS 2449, pp.9-16, 2002.  
 Ribo et al: Hybrid Tracking for Outdoor AR Applications, IEEE Computer Graphics and Applications 2002, pp 54-63  
 Chen, Pinz: Structure and Motion by Fusion of Inertial and Vision-Based Tracking, Submitted to ICPR 2004  
 Stock, Pinz: Object-centered Feature Selection for weakly-Unsupervised Object Categorization, Submitted to ICPR 2004