A Multi-Camera 6-DOF Pose Tracker

College of Computing, Georgia Institute of Technology, Atlanta, GA
Sarah Tariq and Frank Dellaert

Abstract

Most of the work in head-pose tracking has concentrated on single-camera systems with a relatively small field of view which have limited accuracy because features are only observed in a single viewing direction. We present a multi-camera pose tracker that handles an arbitrary configuration of cameras rigidly fixed to the observer's head. By using multiple cameras, we increase the robustness and accuracy by which a 6-DOF pose is tracked. However, in a multi-camera rig setting, earlier methods for determining the unknown pose from three world-to-camera correspondences are no longer applicable. We present a RANSAC [2] based method that handles multi-camera rigs by using a fast non-linear minimization step in each RANSAC round.

1 Multi-Camera Pose Tracking

In a multi-camera rig setting, earlier methods for determining the unknown pose from three world-to-camera correspondences are no longer applicable, as they all assume a common center of projection [2, 5, 4].

We assume a model whereby n landmarks $\{P_j\}_{j=1}^n$ are observed by a multi-camera rig consisting of m cameras with calibration matrices $K_R \overset{\Delta}{=} \{(K_i, R_i, t_i)\}_{i=1}^m$, as illustrated in Figure 1, yielding n measurements $\{(i_j, p_j)\}_{j=1}^n$. Hence, given the global pose (R,t) of the entire rig in a given reference frame, we obtain the following measurement equations for each 3D to 2D correspondence (P,i,p):

$$p = \Pi_i(K_i, R_i(R(P-t)-t_i) + n_i$$

where n_i is a 2D noise vector, and the projection Π is

$$\Pi([X,Y,Z]^T) = [f_x x + sy + u_0, f_y y + v_0]^T$$

Given a list of n correspondences $\{(P_j, i_j, p_j)\}_{j=1}^n$, we estimate the rig pose (R,t) by maximum a posteriori (MAP) estimation:

$$(R,t)^* \quad = \quad \underset{R,t}{\operatorname{argmax}} \left\{ P(R,t) \prod_{j=1}^{n} P(P_j, i_j, p_j | R,t) \right\} (1)$$

where we applied Bayes law and assumed conditional independence of all measurements p_j given the rig pose. The prior $P(R,t)$ can be derived from the previous time step.

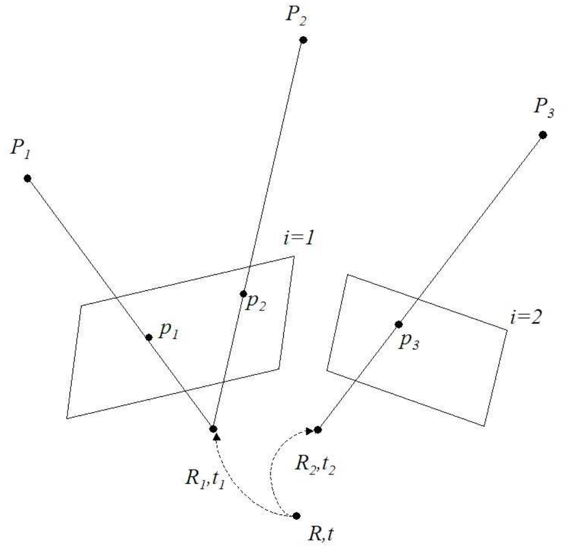

Figure 1. Multi-camera rig example.

In the common case of assumed Gaussian noise, (1) becomes the following non-linear minimization problem

$$(R,t)^* = \underset{R,t}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{j=1}^{n} J(P_j, i_j, p_j) - \log P(R,t) \right\} \quad (2)$$

where $J(P_j, i_j, p_j)$ is the objective function contribution resulting from the j^{th} correspondence, given by

$$J(P,i,p) \overset{\Delta}{=} \|p - \Pi_i(K_i, R_i(R(P-t)-t_i)\|^2_{\Sigma_i} \quad (3)$$

Here $\|\mu - x\|^2_\Sigma$ in (3) is the squared *Mahalanobis distance*. In our implementation, minimizing the objective function (2) is implemented using sparse Levenberg-Marquardt.

In a tracking context, we then use RANSAC to obtain a robust pose estimate using the machinery in Section 1 as a subroutine. At each step, we assume that a number of putative 3D to 2D correspondences $\{(P_j, i_j, p_j)\}_{j=1}^N$ can be obtained, with $N \gg 3$. In Section 2 below we present one way to do this, but any method will do. We then use RANSAC [2] to obtain a set of inlier correspondences. Briefly, we randomly select minimal sets of 3 correspondences, obtain the MAP pose using (2), and check for support among the other inliers. We use an adaptive threshold version of RANSAC to automatically determine the number of RANSAC rounds needed, see e.g. Hartley and Zisserman [3] for a thorough exposition. As a final step, the basis set of correspondences with the highest support is then used with its inlier support to refine the MAP pose estimate.

2 A Markerless Multi-Camera Tracker

Based on this method we implemented a markerless tracking system using affine invariant features [1]. We implemented the entire run-time pipeline from images to pose, but were not yet able to test the system in real environments.

In the surveying phase we detect affine invariant features in the environment using the method outlined in [1] and log them in a database. The location estimation for the features can be done using structure from motion approaches [3]. To compress the database and enable faster comparisons between features we perform principal component analysis on the descriptors, keeping only the first 20 eigenvectors.

In tracking we estimate at each frame the absolute pose of the rig relative to the environment. To estimate the pose we first detect affine invariant features in the images from the rig and then project them into the eigenspace of the database. Putative correspondences between the current affine invariant features and those in the database are computed by finding for each feature in the current images the closest feature in the database. All correspondences with an error larger than a predefined threshold are discarded.

3 Results and Discussion

To demonstrate the quality of the proposed system we conducted experiments in a synthetic environment consisting of texture mapped planes. We used real human motion capture data captured at a rate of 120 frames per second. All experiments were done on an Intel Pentium 4 machine running at 2.80 GHz. Below we present results from one of the sequences, SS1, a large sequence (3386 frames) with complicated motion and relatively large out of plane rotations.

In Figure 2 we show the translational and rotational average absolute errors of the estimated path from the ground truth graphically. Both translational and rotational errors decrease substantially as the number of cameras is increased, and this happened consistently in all our experiments using various types of of mocap sequences. Figure 3 (a) and (b) plot a section of the time-series of both translation along the X-axis and the tilt, for 1 camera and 4 cameras, respectively. From the figures one can see that a considerable number of catastrophic failures occur with just one camera. These results convincingly demonstrate the advantage of using a multi-camera rig tracker over a single, limited field of view camera.

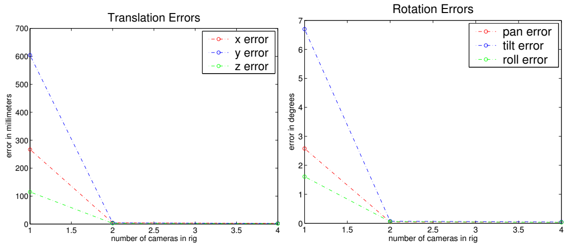

Figure 2. Average deviation from ground truth for a varying number of cameras

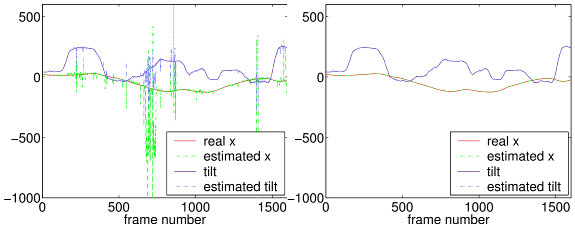

Figure 3. Real and estimated translations along x and real and estimated tilt.

In conclusion, we have shown that, in the context of the markerless tracking system we developed, our system outperforms single-camera systems by a wide margin. We tested the system in software on realistic image sequences, using motion capture data to guarantee realistic motion. To validate in real-time on real image sequences we are developing a FPGA-based miniature camera rig that will be able to perform the detection of affine invariant features in real time.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 774–781, 2000.

[2] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, 24:381–395, 1981.

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

[4] L. Quan and Z.-D. Lan. Linear n-point camera pose determination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(8):774–780, 1999.

[5] R.Haralick, C.Lee, K.Ottenberg, and M.Noelle. Review and analysis of solutions to the three point perspective pose estimation problem. *Intl. J. of Computer Vision*, 13(3):331–356, 1994.