3D Motion Estimation & Tracking

Suya You
Integrated Media Systems Center
Computer Science Department
University of Southern California

June 2004
Tracking Is Major Obstacle

- Sensor pose tracking
  - Where am I, and where am I looking?
  - 6DOF measurement
  - High precision and robust
  - Unprepared environments

- “One of the areas that has seen insufficient innovation in the past decade, position and orientation tracking…” (National research council report on modeling and simulation, 1997)
A long research history in computer vision, navigation…

- **Active approach**
  Powered sensor - emitter in environment
  *Ultrasonic, laser range finder, optical sensor…*

- **Passive approach**
  Computer *vision, compass, inertial gyro, accelerometer…*

- **Hybrid approach**
  Combining the characteristics of single approach to compensate for the weaknesses in each approach
  *Magnetic/vision, inertial/vision, inertial/ultrasonic…*
Explore various tracking technologies

- **Indoor environment**
  - Mark based vision approach
    - Multi-ring color landmark system - scalable
    - Cluster recognition - large workspace
    - Square landmark system – robustness, real-time

- **Outdoor unprepared environment**
  - Structure from motion (SFM)
  - Model based approach
  - Auto-calibration of “new feature (point, line)”
  - Hybrid vision, Inertial, and GPS sensors
  - Panoramic imaging sensor
  - Sensor fusion and closed-loop stabilization
    - Hybrid inertial/GPS/vision tracking
    - Complementary filters and robust estimation
Approach: Wide Area Tracking With Panoramic Images

- Theory issues about the traditional vision system
  - Limited working volume size
  - Large motion
  - Partial occlusion
  - Motion ambiguities (pure translation and rotation)
Benefits Using Panoramic Imaging

- Wide FOV ensures:
  - A sufficient number of features for tracking
  - Less partial occlusion
- Accurate estimates for large motion
  - Provides sufficient information for distinguishing motion ambiguities (pure translation and rotation)
Technical Approach

- Use multiple panoramic images
  - Robust Natural Feature Tracking
- Use 2D feature motion to compute 5DOF camera motion
  - Recursive Rotation Factorization (RRF)
- Use multiple 5DOF motion estimates to compute 6DOF pose
  - Large Motion Pose Estimation (LMPE)
5DOF Motion Estimation

- Translation estimate is crucial
  - Large translation improves pose estimate
  - Parallax effects

- Existing methods are hard to used
  - Developed for perspective imaging (e.g. 8-points method)
  - Insufficient correspondence information

- Proposed approach
  - Panoramic imaging provides sufficient feature information
  - Recursive Rotation Factorization (RRF)
    - Estimate translation and rotation concurrently and incrementally
    - Motion estimates are incrementally improved with each feature providing partial information about the camera motion
▪ RRF provides the iteration control to achieve accurate translation estimates for the general mixtures of large translation and rotation motions

▪ IKEF performs motion estimates incrementally with each feature providing partial information about the camera motion

Large 5DOF Motion estimation procedure
6DOF Pose Estimation

- 6DOF pose directly from two 5DOF motion estimates (orientation and direction of translation)
  - Two reference images
  - 2D-2D tracking

- No dependence on scene calibration
  - Calibrated features
  - 2D-3D correspondences
LMPE Approach Overview

Reference image (1) → 5DOF Motion estimation procedure → Orientation estimate

Current image → 5DOF Motion estimation procedure → Translation estimate

Reference image (2) → 5DOF Motion estimation procedure → Orientation estimate

Translation estimate → 6DOF pose estimate

5DOF estimation (RRF)

6DOF estimation (LMPE)
Results: Lab Test Data

- Lab test based on manual calibrated grid positions (5m x 5m)
- HiBall ceiling tracker was used as “ground-truth” reference
Results: Lab Test Data

- Charts compare LMPE vs. traditional 3D-2D method
  - Green lines indicate the LMPE estimates
  - Red lines indicate the 3D-2D estimates
  - Black lines indicates the measured position
Results: Outdoor Scene

- Uncalibrated outdoor environments
- Wide tracking range tested to 50 meters
- GPS reading as “ground truth”
Natural features are detected/tracked automatically for pose estimate.

33 features are detected in the 1st reference image.

23 features are tracked in the 748th tracked image.
Results: Outdoor Scene (con.)

- Charts show estimated position vs. GPS measurements
  - Green (X), Red(Y), Blue (Z), and Black indicates GPS measures
  - 2 reference image are used at each tracking frame
  - 25cm tracking difference compared with GPS reading
Estimated camera pose is used for visual navigation
Hybrid Tracking Technology

- Inertial sensor
  
  Advantages:
  - Self-contained, and sourceless
  - High sample rate
  
  Disadvantages:
  - Need sensor signal integration for attitude computation that results in accumulated *drift* with elapsed time

- Vision sensor
  
  Advantages:
  - Tracking measurements made from the viewing position often minimize the visual alignment error
  
  Disadvantages:
  - Lack of robustness, end-to-end system delay, and high computational expense
Purpose & Technical Challenges

- Develop hybrid vision/inertial sensors to produce *stable* motion tracking
  - Combine the characteristics of vision tracking with the gyro sensors to compensate for the weaknesses in each component
  - Stabilize camera pose in large-area environments
  - Reduce system delay

- Robust measurement and fusion
  - Varying sensor data availability
  - Varying certainty of measurements
  - Fusion models and algorithms
Technical Approaches

- Hybrid vision/gyro 3DOF orientation tracking
  - 3D DOF orientation
  - Natural feature tracking
  - SFM (structure from motion) vision algorithm

- Hybrid vision/gyro 6DOF pose tracking
  - 6DOF camera pose
  - Landmark based tracking
  - Complementary motion estimate filter
    (Suya Y and Ulrich N, “Fusion of Vision and Gyro Tracking for Robust Augmented Reality Registration”, IEEE Virtual Reality, March 2001)
Designed Inertial Gyro Sensor

- Three orthogonal gyroscopes (max sense range: +/-500 deg/sec)
- AD/gyro sample at 1KHz
Error Sensitivity of Inertial Tracking

- **Static error analysis:**
  - Keep the sensor still
  - Sample the angular rate at 1kHz
  - Integration of the angular rates at 30Hz
  - Third-order term integration approximating
  - The measured drift rate is about 0.7 degree/min

- **Dynamic error analysis**
  - The 3DOF gyro sensor is rigidly attached to video camera
  - Coordinate transformation produces three orientation measurements (Yaw, Pitch, and Roll) with respect to the initial orientation
  - Select several image features for evaluation
  - Compare observed feature positions to projected positions reported by the gyroscopes
Dynamic Error Analysis of Inertial Tracking

Average pixel differences between tracked features and back-projected features of gyroscope sensors
Designed Hybrid Vision/gyro Tracker

- **Hardware:**
  - CCD video camera
  - Three gyroscopes
  - Data sample/filer devices

- **Software:**
  - Gyro sample lib
  - Gyro algorithm lib
  - Video capture lib
  - Vision tracking lib
  - Fusion algorithm lib
Approach: Hybrid Vision/INS 3DOF Pose Tracking

- **Goal**
  - Combine a nature feature tracking system with gyro sensors to provide 3DOF orientation tracking

- **Approach**
  - The inertial data serves as an aid to the vision tracking by reducing the search space and providing tolerance to interruptions
  - Vision corrects for inertial drift accumulation

- Static calibration
- Pose estimation

Diagram:
- Gyro
- Video
- 2D predictor
- 2D Tracker Vision
- 3D orientation corrector
- Corrected orientation
Static Calibration

- **Camera calibration**
  - Camera intrinsic parameters and the lens distortion parameters

- **Gyroscope calibration**
  - Zero-voltage offset (bias) – averaging readings
  - Scale factors – provided by manufacturer’s test sheets

- **Calibration between inertial and camera frames**
  - The relationship between the camera and inertial coordinates:
    
    \[
    \omega_c = [R_{lc}] \omega_i
    \]
Under perspective projection, the 2D-image motion resulting from camera motion can be written as

\[
\begin{align*}
\dot{x}_u &= \left[ -fV_{cx} + xuV_{cz} + xu y_u \omega_z - f(1 + \frac{x_u^2}{f^2}) \omega_y + y_u \omega_x \right] \\
\dot{y}_u &= \left[ -fV_{cy} + yuV_{cz} + f(1 + \frac{y_u^2}{f^2}) \omega_x - \frac{x_u y_u}{f} \omega_z + x_u \omega_y \right]
\end{align*}
\]

Eliminating the translation term and re-organizing:

\[
\begin{align*}
\dot{x}_u &= \Lambda \left[ R_{IC} \right] \omega_1 \\
\Lambda &= \left[ \begin{array}{ccc}
x_u y_u & -f(1 + \frac{x_u^2}{f^2}) & y_u \\
y_u^2 & x_u y_u & -x_u \\
f(1 + \frac{y_u^2}{f^2}) & \frac{x_u y_u}{f} & -x_u \\
\end{array} \right]
\end{align*}
\]
Pose Estimation

- Visually track a set of nature features
- Inertial gyro provides frame-to-frame predictions of camera motion to increases the robustness and computing efficiency of vision system
- Calibrated camera maps pixel-motion to 3D rotation
- Measured gyro-rotation maps to camera-rotation through a calibration transformation
2D Tracking Prediction

- Suppose $N$ features are detected in the scene. We want automatically track these features as the camera moves.

- The motion positions of these points, due to the related motion (rotation) between the camera (gyro) and the scene:

$$
\mathbf{x}_u^t = \mathbf{x}_u^{t-1} + \Delta \mathbf{x}_u^t
$$

- Use inertial data predicts the motion of image features

$$
\Delta \mathbf{x}_u^t = \Lambda \mathbf{\omega}_c^t
$$

2D Tracking Correction

- The correction refines the predicted image feature positions by doing local searches for the true positions.
- A robust 2D motion tracking approach is used for the correction strategy:
  - Select the “best” features for tracking.
  - Closed-loop Integrated architecture (detection, tracking, and verification).
  - Multi-resolution strategy.

3D Tracking Correction

\[ \omega_I = \omega_c + \Delta\omega \]

\( \Delta\omega \) is the gyro drift that we want to estimate and correct

- Find the relationship between the gyro error and resulted 2D error of image velocity:

\[ \dot{x}^I - \dot{x}^C = \Lambda \cdot \Delta\omega \]

- So, to find the appropriate inertial drift \( \Delta\omega \), so that the motion residual \( \|\dot{x}^I - \dot{x}^C\| \rightarrow \text{min} \) reaches minimum

\[ \Delta\omega = \Lambda^{-1} \cdot (\dot{x}^I - \dot{x}^C) \]
System Configuration

- Video Recorder
- Sony XC999 video camera
- 3 GyroChip II rate gyros
- V-Cap optical see-through HMD
- 16-bit A/D
- 200 MHz Pentium PC
- VGA video
Results – USC Campus

- Average error is 5.1 pixels (~0.6 degree of motion angle)
Results – HRL Laboratory

Cooperated with HRL

- Average error is 4.27 pixels (~0.4 degree of motion angle)
Approach:
Hybrid Vision/INS 6DOF Pose Tracking

- **Goal**
  - Stabilize 6DOF pose tracking

- **Motivation**
  - Vision sensors (cameras) nominally sample at video rates (30Hz), which are most appropriate for measuring *low-frequency* pose variations. Rapid or abrupt camera rotations or motions can cause vision tracking failures or instabilities.
  - Inertial sensors can be sampled at very high rates (~1KHz), which makes them suitable for sensing the rapid motions that create *high-frequency* pose variations, and provide a frame to frame prediction of camera orientation.
  - A *complementary* motion estimate filter can fuse those different sample rates sensors.
Motion Model and System Dynamics

- Camera motion at time $t_i$

$$M_i = (r_i, v_i, \psi_i, \omega_i)$$

- Constant velocity system dynamics

$$M_{i+1} = \begin{bmatrix} r_{i+1} \\ v_{i+1} \\ \psi_{i+1} \\ \omega_{i+1} \end{bmatrix} = \begin{bmatrix} r_{i} + v_{i} \Delta T \\ v_{i} + n_v \\ \psi_{i} + \omega \Delta T \\ \omega_{i} + n_{\omega} \end{bmatrix}$$
Complementary motion estimate filter accommodates the different sample-rate sensors and reduces the end-to-end system latency

- Two independent correction channels share a common prediction module to handle incomplete information measurements
**Fusion Filter Implementation**

- **EKF** (could be others – *what’s the problem?*) is used for the filter implementation, which can be treated as two parallel EKF banks sharing one common state prediction module.

- **Image measurement**

  \[
  \begin{bmatrix}
  P_{x,i}^k \\
  P_{y,i}^k
  \end{bmatrix} = 
  \begin{bmatrix}
  f \frac{R_1(\psi_i)}{R_3(\psi_i)} \cdot [P_k - r_i] + n_x \\
  f \frac{R_2(\psi_i)}{R_3(\psi_i)} \cdot [P_k - r_i] + n_y
  \end{bmatrix}
  \]

- **Inertial gyro measurement**

  \[
  \omega_i = W^{-1}(\theta_i) \dot{\theta}_i
  \]
System Implementation

- Fiducial system (detection and recognition)
- Gyro system (calibration and measurement)
- Fusion system (fusion and attitude computation)
- CCD video camera
- Three gyroscopes
- Data sample devices
- Calibrated landmark array for indoor experiments
Results: Detection & Recognition

- 50 trained patterns
- Camera undergoes arbitrary 6DOF motions
- 96% recognition rate
- 0.87-pixels average RMS detection accuracy
Results: Pose Estimate

- Camera undergoes arbitrary 6DOF motions, viewing the calibrated board
- Camera pose is continually computed by the fusion filter in real time
- Use back-projection to measure tracking accuracy
  - Green cross-hairs: detected features
  - Red cross-hairs: back-projected features based tracked camera pose
Results: Performance

- Dynamic accuracy: 2.18-pixels average error, and 9.93 pixels max error
- Static error: less 1.3 pixels
- 23 fps (on a 450 MHz Pentium III)
Approach:
Hybrid Vision/INS/GPS 6DOF Pose Tracking

- Goal
  Tracking and stabilizing camera 6DOF pose estimate in large-area outdoor environments
    - Easily acquired a priori knowledge – scene models
    - Robust – self correcting and initializing
    - Real-time operation
Designed Hybrid Vision/INS/GPS Tracker

- Portable tracking package
  - DGPS (Z-Sensor base/mobile from Ashtech)
  - INS (IS300 from Intersense)
  - Stereo camera head (MEGA-D from Videre Design)
  - Real-time data acquisition and display
    - GPS: ~1Hz
    - INS: ~150Hz
    - Video: ~30Hz
  Synchronize & fuse at 30Hz video rate
Tracking Needs to be Stabilized

- Accuracy is not enough
  - Error is easily visible and undesirable
  - DGPS: ~10cm (open space)
  - INS: ~1 degree (static) - One degree of orientation error results in about 11-pixels of alignment error in the image plane
Approach Overview

- **INS predicts feature motion**
- **Low frame-rate (~1 Hz) video snapshots correct for gyro, translation, and calibration errors**
- **Absolute pose from 3D model DB**

**Integrate INS**
- 3D rate gyro

**2D motion prediction**
- calibration
- synchronization

**Feature search**
- lines, ...

**2D / 3D correspondence**
- model DB

**6DOF Pose & Structure Estimate**

- **KF 6DOF pose estimation**
  - Pose for the current frame
  - 2D features in current frame
  - 3D feature database
  - Relative orientation from gyro
  - Prior video image
  - Current video image
  - Features detected in prior frame

**2D tracking**
System Components

- 2D/3D image feature detection and tracking
  - Edge/line features prominent in urban environments
- Fusion of GPS, INS, and vision sensors
  - Portable GPS/GYRO/VISION tracking system
- Low frame-rate (~1 Hz) vision tracking
  - Correspondence over large motions
- Model based pose estimate
  - 6DOF pose recovering
  - Line/edge based approach
  - Need accurate scene 3D model
    - Rapid scene model construction - aerial LiDAR data
Predict 2D feature motions

- Pure rotation produces uniform feature motion – regardless of scene depth
- Assumption – rotation is dominant cause of feature motion
  - Valid for short periods (sec’s) where features are relatively far and motion is at human movement rates

Lines/edge features

- More robust than point features (noise, occlusion…)
- Physical meaning – corresponds to 3D model (edge, window…)


2D Line Detection and Tracking

- For a line detected on the last frame, predict its end points (blue) on the current frame to get a predicted 2D line.
- For points along the predicted line (black), find corresponding points on the last frame detected line for gradient measures.
- Find peak gradients for the predicted points along an interval perpendicular to the predicted line and use gradient match to determine valid line points.
- Use all the valid points on the to fit a tracked current frame line (blue).
- Note: For vision-only tracking, the prior frame line is used as the predicted line for the new frame.
### 2D Line Detection and Tracking (con.)

#### Detection routine

1. **Intensity Normalization**
   - Multi-dimension vector \((X, L, a, I)\)
   - Parametric form \((AX+BY+C = 0)\)

2. **Feature detection**
   - Line trace (chain coding)

3. **Line segments representation**
   - Multi-dimension vector \((X, L, a, I)\)
   - Parametric form \((AX+BY+C = 0)\)

#### Tracking routine

1. **Line segments representation**
   - Two end points \((X)\)
   - Length \((L)\)
   - Angle \((a)\)
   - Line intensity \((I)\)

2. **Tracking search (vector matching)**
   - \[ S = w_1 \Delta X + w_2 \Delta L + w_3 \Delta a + w_4 \Delta I \]

3. **Tracked line segments**
   - Multi-dimension vector \((X, L, a, I)\)
   - Parametric form \((AX+BY+C = 0)\)
Extension to Window Detection and Tracking

- **Window pattern**
  - Four point polygon
  - Allow perspective deformation
  - Four corners are visible

<table>
<thead>
<tr>
<th>Intensity Normalization</th>
<th>Coarse detection</th>
<th>Projection compensation</th>
<th>Fine detection</th>
<th>Tracking search</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 2D image gradient</td>
<td>• Motion prediction</td>
<td>• Perspective-imaging model to compensate for geometric deformations</td>
<td>• Fitting the extracted candidates to the defined models</td>
<td>• Four corner SSD matching</td>
</tr>
<tr>
<td>• Light compensation</td>
<td>• Find the high contrast regions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data base

Feature recognition

• PCA based recognition

(VR 2001)
Search for features and correspondence

- Line extension/connection tests – prevent shrinkage/loss/fragmentation
  - constraints and heuristics for sanity checks
  - 2D direction (vanishing points) – new line detection
  - 3D model plane – model feature addition/correction
  - Area features (windows, doors, ...) with correlation/recognition signatures
Suppose \((T_1, R_1)\) is the estimated pose for the last frame, and \(\dot{R}_g\) is the integrated orientation change measured by the gyro.

- \((T_1, \dot{R}_g \cdot R_1)\) is the current frame pose prediction.
- In vision-only case, \((T_1, R_1)\) is the current frame prediction.

\((T_1, \dot{R}_g \cdot R_1)\) is a better prediction that leads to faster more robust pose EKF convergence.

- especially when the relative orientation change is large.
Real-time Vision Computation

- **Multi-thread structure**
  - Video acquisition (capture, color-grey convert…)
  - 2D processing (feature detection and tracking)
  - 3D processing (pose and structure)
  - File I/O, display (model load, rendering)

- **Fast image processing**
  - Most time-consume part is 2D image processing (gradient, convolution)
  - 320x240 (whole size), 7x7xN, 320x120 (half size), …
  - Image is adaptively segmented and processed only the half (or less)

- **Occlusion (feature disappear, out of view…)**
  - Minimum number check to maintain a rough constant number
  - Re-detect and update new features
  - Estimated 3D structure is fed back for a smooth motion
Summary

- Explored the complementary characteristics of single tracking approach to compensate for the weaknesses in each approach
  - Combine vision tracking with inertial sensors to produce stable motion tracking
- Proposed approaches from different point of view to achieve robust measurement and fusion
  - Varying sensor data availability
  - Varying certainty of measurements
  - Fusion models and algorithms
  - System delay