Feature Tracking and Motion Factorization from Monocular Video

Lakulish Antani

December 1, 2006

Outline

Goal

To develop the components of a motion-capture system based on input from a single camera.

Components of a System

- Main components of a single-camera mocap system:
  - Video capture (input)
  - Feature tracking
  - Structure-from-motion
  - Visualization
- We look at tracking and structure-from-motion (using factorization)

Problem Statement

- INPUT: Image sequence $I(\mathbf{x}, t)$
- INPUT: Set of "feature" points in frame t_0
- OUTPUT: Positions of feature points in frame $t_0 + 1$

KLT Algorithm

- Assume the model:

$$I(\mathbf{x}, t) = I(\mathbf{x} + \mathbf{d}, t + \Delta t)$$

- Consider rectangular "windows" around each feature point
- Minimize the SSD error of pixel intensities in the window:

$$\epsilon = \int_W [I(\mathbf{x} + \mathbf{d}, t + \Delta t) - I(\mathbf{x}, t)]^2 w(\mathbf{x})$$

- $w(\mathbf{x})$ is a weighting function, usually 1

KLT Algorithm

- Using first-order Taylor expansion, and setting derivative w.r.t. **d** to 0:

$$\mathbf{G}\mathbf{d} = \mathbf{e}$$

- $\hat{\mathbf{d}} = \mathbf{G}^{-1}\mathbf{e}$ is returned as the optimal value of displacement

Details

$$\begin{aligned} \mathbf{G} &= \sum_{W} \begin{bmatrix} g_1^2 & g_1 g_2 \\ g_1 g_2 & g_2^2 \end{bmatrix} \\ \mathbf{e} &= \sum_{W} h \begin{bmatrix} g_1 & g_2 \end{bmatrix}^T \\ g_i &= \frac{\partial I}{\partial x_i}, i = 1, 2 \end{aligned}$$

Smoothing the Displacements

- KLT ignores coherence between features on the same rigid body
- Transform (componentwise) the **d** field to get a "smooth" field of vectors **v**
- Use a weighted average of **d** and **v** as the final displacement
- Smoothing is done using Hierarchical Radial Basis Function Networks

HRBF Networks

- Represent $f(x)$ as a weighted sum of Gaussians
- Gaussians are grouped into *levels*, which determine their density and variance:
  - Gaussians at the same level have the same variance
  - Centers are on points in a regular grid

Details

$$f(\mathbf{x}) = \sum_{i=1}^{L} a_i(\mathbf{x})$$

$$a_i(\mathbf{x}) = \sum_{j=1}^{M_i} w_{i,j} g(\mathbf{x} - \mathbf{x}_j; \sigma_i)$$

$$g(\mathbf{x}; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\|\mathbf{x}\|^2}{\sigma^2}}$$

HRBF Networks

- Begin by computing $a_0(x)$, the approximation at level 0
- At subsequent levels, fit points of the form $\{\mathbf{x}_i, \Delta^j y_i\}$ where

$$\Delta^j y_i = f(\mathbf{x}_i) - \sum_{k=1}^{j-1} a_k(\mathbf{x}_i)$$

- Stop when the residue (Δy_i^j) is "small" enough

HRBF Networks

- Weights are computed by taking a local weighted mean:

$$w_{i,j} = \frac{\sum_{\mathbf{x} \in N_j} s_i(\mathbf{x}) g(\mathbf{x} - \mathbf{x}_j; \sigma_i)}{\sum_{\mathbf{x} \in N_j} g(\mathbf{x} - \mathbf{x}_j; \sigma_i)}$$

where

$$s_i(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & i = 0 \\ f(\mathbf{x}) - \sum_{k=1}^{i-1} a_k(\mathbf{x}) & i > 0 \end{cases}$$

- Gaussians are inserted at levels beyond 0 only if there is sufficient residual error:

$$\sum_j \frac{\|\Delta^k y_j\|}{\|N_j\|} > \epsilon$$

Experimental Results

Standard KLT

- Select and track "best" 50 features
- Attempt to reselect lost features

Experimental Results

KLT with HRBF Smoothing

- Select and track "best" 50 features
- Attempt to reselect lost features
- Use $\epsilon = 0.0001$ and $w = \frac{1}{4}$

Problem Statement

- INPUT: P points tracked over F frames, $\{(u_{fp}, v_{fp})\}$, as a $2F \times P$ matrix:

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$$

  $\mathbf{W}$ is *registered* (giving $\hat{\mathbf{W}}$) by subtracting from each entry the mean of the entries in its row

- OUTPUT: $2F \times 3$ matrix $\mathbf{R}$, representing motion; and $3 \times P$ matrix $\mathbf{S}$, representing structure

Rigid Body Assumptions

- Assume there is 1 rigid body, and an orthographic projection model
- In frame f with camera axes $\mathbf{i}_f$ and $\mathbf{j}_f$, for a point $\mathbf{s}_p$:

$$\begin{aligned} u_{fp} &= \mathbf{i}_f^T \mathbf{s}_p \\ v_{fp} &= \mathbf{j}_f^T \mathbf{s}_p \end{aligned}$$

- In other words:

$$\hat{\mathbf{W}} = \mathbf{RS}$$

Tomasi-Kanade: SVD

- Note that $rank(\hat{\mathbf{W}}) \leq 3$
- The Singular Value Decomposition of $\hat{\mathbf{W}}$:

$$\hat{\mathbf{W}} = \mathbf{O_1 \Sigma O_2}$$

- We define:

$$\begin{aligned} \hat{\mathbf{R}} &= \mathbf{O_1 \Sigma^{\frac{1}{2}}} \\ \hat{\mathbf{S}} &= \mathbf{\Sigma^{\frac{1}{2}} O_2} \end{aligned}$$

giving

$$\hat{\mathbf{W}} = \hat{\mathbf{R}}\hat{\mathbf{S}}$$

Tomasi-Kanade: Constraints

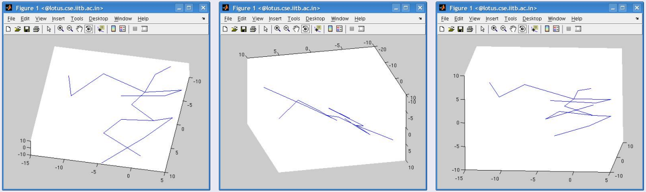
- Previous decomposition is not unique
- For any invertible $\mathbf{G}$,

$$(\hat{\mathbf{R}}\mathbf{G})(\mathbf{G}^{-1}\hat{\mathbf{S}}) = \hat{\mathbf{R}}\hat{\mathbf{S}} = \hat{\mathbf{W}}$$

- Impose the constraints:
  - Rows of $\mathbf{R}$ are unit vectors
  - The first F rows of $\mathbf{R}$ are orthogonal to the second F
- Solve for $\mathbf{G}$ and use $\mathbf{R} = \hat{\mathbf{R}}\mathbf{G}$ and $\mathbf{S} = \mathbf{G}^{-1}\hat{\mathbf{S}}$

Experimental Results

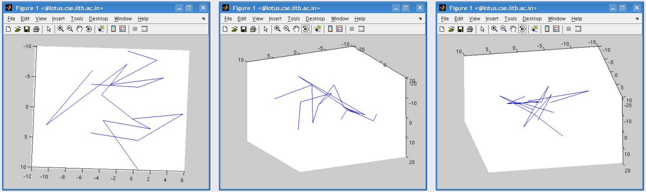
Input from Standard KLT



Key points from the factorization output, joined by lines to form a skeleton.

Experimental Results

Input from KLT with HRBF Smoothing



Key points from the factorization output, joined by lines to form a skeleton.

Conclusions

- HRBF-enhanced tracker needs fine-tuning before it can perform significantly better than standard KLT
- Both trackers have trouble with occlusions
- Tomasi-Kanade works well with the trackers
- A non-rigid factorization component is needed

Further Work

- Implementing a non-rigid factorization component
- Further enhancements to the tracker
- Packaging these components as a complete system, along with visualization

References

Ferrari, S., Maggioni, M., and Borghese, N. A.
Multiscale approximation with hierarchical radial basis function networks.
In *IEEE Transactions on Neural Networks, Vol. 15. No. 1* (2004).

Gonzalez, J. J., Lim, I. S., Fua, P., and Thalmann, D.
Robust tracking and segmentation of human motion in an image sequence.
In *ICASSP, Vol. 3* (2003).

Lucas, B., and Kanade, T.
An iterative image registration technique with an application to stereo vision.
In *Proceedings of Image Understanding Workshop* (1981), pp. 121–130.

Shi, J., and Tomasi, C.
Good features to track.
In *IEEE Conference on Computer Vision and Pattern Recognition* (1994).

Tomasi, C., and Kanade, T.
Detection and tracking of point features.
In *CMU Technical Report CMU-91-132* (1991).

Tomasi, C., and Kanade, T.
Shape and Motion from Image Streams: a Factorization Method, Full Report on the Orthographic Case.
Tech. Rep. CMU-CS-92-104, March 1992.

Ullman, S.
The Interpretation of Visual Motion.
MIT Press, 1979.