Bayesian Reconstruction of 3D Human Motion from Single-Camera Video

Nicholas R. Howe
Cornell University

Michael E. Leventon
MIT

William T. Freeman
Mitsubishi Electric Research Labs
Problem Background

- 2D video offers limited clues about actual 3D motion.
- Humans interpret 2D video easily.
- **Goal**: Reliable 3D reconstructions from standard single-camera input.
Research Progress

- Multi-camera trackers available:
  1996: Gavrila & Davis; Kakadiaris & Metaxas
- Potential single-camera trackers:
  1995: Goncalves et. al.
  1997: Hunter, Kelly & Jain; Wachter & Nagel
  1998: Morris & Rehg; Bregler & Malik
- Previous work: treated as measurement problem, not inference problem.
Challenges

• Single camera
  ⇒ 3D ambiguity
  (underconstrained problem)
  ⇒ Foreshortening
  ⇒ Self-occlusion

• Unmarked video (no tags)
  ⇒ Appearance changes
  ⇒ Shadowing
  ⇒ Clothing wrinkles
Overview of Approach

• Two stages to tracking, each challenging:

2D Tracking  →  3D Reconstruction
2D Tracking

- Predict 2D Pose, Model
- Refine 2D Pose
- Compare with Image
- 2D Pose + Model = Rendering

• Repeat for each frame.
2D Tracking Details

• Pose for first frame is given.
• Model derived from past frames.
  – We use “part map” models.
• For each frame, begin at low resolution and refine.
• Rendering must account for self-occlusions. (need 3D feedback!)
Occlusion

- Must compute hidden pixels given pose.
- Only visible pixels matched with image.
- Model for hidden regions not updated.
2D Tracking Performance

- Simple example, no occlusion:

Lines show tracked limb positions.
3D Reconstruction

- Motion divided into short movements, informally called *snippets*. (11 frames long)
- Assign probability to 3D snippets by analyzing knowledge base.
- Each snippet of 2D observations is matched to the most likely 3D motion.
- Resulting snippets are stitched together to reconstruct complete movement.
Learning Priors on Human Motion

- Collect known 3D motions, form snippets.
- Group similar movements, assemble matrix.
- SVD gives Gaussian probability cloud that generalizes to similar movements.
Posterior Probability

- Bayes’ Law gives probability of 3D snippet given the 2D observations:

\[ P(\text{snip} \mid \text{obs}) = k \ P(\text{obs} \mid \text{snip}) \ P(\text{snip}) \]

- Training database gives prior, \( P(\text{snip}) \).
- Assume normal distribution of tracking errors to get likelihood, \( P(\text{obs} \mid \text{snip}) \).
Posterior Probability (cont.)

- Posterior is a mixture of multivariate Gaussian.
  \[ P(\tilde{x}, \theta, s, \tilde{v}) = k_1 \left( e^{-\|\tilde{y} - Y_{\theta,s,\tilde{v}}(\tilde{x})\|^2/(2\sigma^2)} \right) \left( \sum_{j=1}^{m} k\pi_j e^{-\tilde{\alpha}_{\tilde{x},j}^T \tilde{\alpha}_{\tilde{x},j}} \right) \]

- Take negative log and minimize to find solution with MAP probability.
- Good solution can be found using off-the-shelf numerics package.
Stitching

- Snippets overlap by 5 frames.
- Use weighted mean of overlapping snippets.
Sample Results: Test Data

- Test on known 3D data:

  Observation

  Reconstruction

  Comparison
Sample Results: Test Data

• Results on wave clip shown earlier:
Sample Results: Real Footage

- Can reconstruct even imperfect tracking:
Conclusion

• Treat 3D estimation from 2D video as an inference problem.
• Need to improve models
  – Body appearance $\Rightarrow$ better rendering/tracking
  – Motion $\Rightarrow$ better reconstruction
• Reliable single camera 3D reconstruction is within our grasp.
Final Video

(Hand-tracked points, automatic reconstruction)
2D Tracking Equation

- Must find pose parameters $\beta$ that minimize matching energy:

$$E(\beta) = \sum_{b \in \text{Body Parts}} \sum_{p \in \text{Points}(b)} \left( \text{Visible}(b, p, \beta) \left[ I_{\text{Model}}(p) - I_{\text{Image}}(\text{Project}(p, \beta)) \right] + E_o(b) \right)$$

Accounts for self-occlusion

Projection of model point into image.

Additional constraints (joints, limb lengths, etc.)
2D Tracking Performance

- Simple example, no occlusion:
Sample Results: Test Data

- Test on known 3D data:
Sample Results: Test Data

- Results on wave clip shown earlier:
Sample Results: Real Footage

- Can reconstruct even imperfect tracking: