Capturing and Understanding Hand Gestures for 3D Tele-presence

Junsong Yuan

School of Electrical & Electronic Engineering and BeingThere Centre

NTU, Singapore
Acknowledgement

BeingThere project 4: Autonomous Virtual Humans and Social Robot for Tele-presence

Hui Liang
Research associate
BeingThere

Daniel Thalmann
Professor
BeingThere and IMI

Nadia Thalmann
Co-director
BeingThere and IMI

NANYANG TECHNOLOGICAL UNIVERSITY
Why using Hands in Tele-presence?

- User experience is critical for successful tele-presence applications
  - Natural and immersive interaction experience
  - Seamless transition between reality and virtual worlds
  - From tele-presence to smart-presence
Random Forest based Body and Hand Pose Analysis

- Regression Forest (RF) + Spatial Voting has proven effective for articulated pose estimation and gesture recognition
  - Ability to handle large intra-class variations
  - Robust to partial occlusion
  - Fast and accurate, real-time or even better


Problems and Challenges

- Gap between training and testing objectives in RF + spatial-voting frameworks
- Multimodal distribution of the pose votes aggregated from spatially-distributed voting pixels
- Articulated hand pose estimation in RGB images is still an unsolved problem
Our case studies

- Spatially-optimized random forest to improve spatial-voting based 6D palm pose prediction
- Multi-modal prediction fusion to leverage motion constraints in articulated hand pose estimation
- Conditional regression forest for joint articulated pose prediction & distance recovery in RGB images

Please contact Hui Liang if you want to see the real-time demo at BTC
Spatially-Optimized Random Forest for 6D Palm Pose Prediction

Task: track the hand and predict unconstrained 3D palm rotation and translation from depth images
Random Forest based Hand Pose Analysis
Spatially-Optimized Random Forest

- Inconsistent training and testing objectives in conventional spatial-voting frameworks for pose estimation

- **Training objective:** minimizing prediction error for individual spatial-voting elements, e.g. pixels or local patches

- **Testing objective:** minimizing fused prediction error for each test image

- **Solution:** retrain the forest with pre-learned tree structures to minimize per-image prediction error
Hui Liang, Junsong Yuan, and Daniel Thalmann, paper under review
Augmented Reality Application
Virtual Reality Application
Our case studies

- Spatially-optimized random forest to improve spatial-voting based 6D palm pose prediction

- Multi-modal prediction fusion to leverage motion constraints in articulated hand pose estimation

- Conditional regression forest for joint articulated pose prediction & distance recovery in RGB images
Multimodal Prediction Fusion for Articulated Hand Pose Estimation

- **Task**: predict 16 3D joint locations of the hand from single depth images
- **Baseline**: random regression forest + spatial voting
Multimodal Prediction Fusion

• The distribution of the votes via per-pixel prediction with RF tends to be multimodal, e.g. middle fingertip

• How can the hand joint correlations resolve the multimodality of the per-pixel prediction?

• Multimodal Prediction Fusion: fuse the per-pixel predictions with linear subspace constraints
Multimodal Prediction Fusion

Products of Experts to model joint posterior distribution \(( \cdot | \cdot )\) of the entire hand joint set given the input image:

\[
P(\Phi | I_D) \propto \prod_i P(\Phi | p_i) = \prod_i \prod_k P(\phi_k | p_i)^{w_{ik}}
\]

**Key idea:** Optimize \(( \cdot | \cdot )\) with respect to \(w_{ik}\) in a learned low dimensional linear subspace

**Problem reformulation:**

\[
*, * = \arg\max ( \cdot | \cdot ) \text{ s.t. } = \Sigma +
\]

**Can be efficiently solved via the EM algorithm**

Hand gesture recognition
Our case studies

Spatially-optimized random forest to improve spatial-voting based 6D palm pose prediction

Multi-modal prediction fusion to leverage motion constraints in articulated hand pose estimation

Conditional regression forest for joint articulated pose prediction & distance recovery in RGB images
Egocentric Hand Pose Estimation and Distance Recovery using RGB Camera

- **Task:** jointly predict the sixteen 2D joints of the hand and the wrist distance using an ordinary Webcam in egocentric view

- **First attempt to solve both tasks in egocentric RGB images**

- **Method:** conditional regression forest with wrist distance as a hidden variable

- **Method:** semantic context to encode hand part for regression

Hui Liang, Junsong Yuan, and Daniel Thalmann, “Egocentric Hand Pose Estimation and Distance Recovery in a Single RGB Image,” IEEE Conf. on Multimedia Expo (ICME’15), 2015

Hui Liang, Junsong Yuan, and Daniel Thalmann, “Parsing the Hand in Depth Images,” IEEE Trans. on Multimedia (TMM’14), 2014
Conditional RF for Hand Pose Estimation and Distance Recovery

- During training, separate trees are trained for each discrete wrist distance.
- During testing, optimal hand joint positions and wrist distance are jointly inferred.
Thank you!

Questions?