Gesture Interaction with Virtual Humans and Social Robots

Daniel Thalmann
Institute for Media Innovation
Nanyang Technological University
Singapore

Visigrapp 2015, Berlin
Overview

Perception

- Face Recognition
- Hand and body gesture Recognition
- Facial Expression Recognition
- Sound Localization and Classification
- Social State Recognition

Sensor Fusion + World Model

Social Networks

Decision making

Emotion Engine

VH/Robot Control

Internal update of VH/robot actions e.g. action completion or failures

Reaction
Body and Hand Gesture recognition
The Nature of Gesture

• Gestures are expressive, meaningful body motions – i.e., physical movements of the fingers, hands, arms, head, face, or body with the intent to convey information or interact with the environment.

• 3 functional roles of human gesture:
  • **Semiotic** – to communicate meaningful information
  • **Ergotic** – to manipulate environment
  • **Epistemic** – to discover environment through tactile experience.
Gesture recognition

• Process by which gestures made by the user are made known to the system.

• Standard mouse and keyboard actions used for selecting items and issuing commands are gestures: trivial cases.

• While static position (posture, configuration, or pose) is not technically considered as gesture, we will include it in gestures.
Kendon’s gesture continuum

- **Gesticulation** – spontaneous movements of hands and arms that accompany speech
- **Language-like gestures** – gesticulation integrated into spoken utterance, replacing particular spoken word
- **Pantomimes** – gestures that depict objects or actions, with or without accompanying speech
- **Emblems** – familiar gestures such as “V for victory”
- **Sign languages** – Linguistic systems, such as American Sign Language

As list progresses (from left to right in figure)
Gestures: Some Concepts

• Recognition of natural, continuous gestures requires temporally segmenting gestures.

• Automatically segmenting gestures difficult, and often ignored in current systems by requiring a starting position in time and/or space.

• Similar to this: problem of distinguishing intentional gestures from other “random” movements.

• No standard way to do gesture recognition, variety of representations and classification schemes used.
Static and Dynamic Gesture

- Gestures can be static, where user assumes certain pose or configuration, or dynamic, defined by movement.
- McNeill defines 3 phases of dynamic gesture:
  - pre-stroke, stroke, and post-stroke
- Some gestures have both static and dynamic elements, where pose is important in one or more of gesture phases; particularly relevant in sign languages.
- When gestures produced continuously, each gesture affected by gesture that preceded it, and possibly by gesture that follows it.
Steps of Gesture Recognition

1. Sensor Processing
2. Feature Extraction
3. Gesture Classification

Body parameters (positions, angles, velocities...) and uncertainty

Feature space representation

Recognized gestures

Gesture DB
Magnetic sensors

- Source generates low frequency magnetic field detected by sensor.
- Polhemus, Ascension
- Calibration retaining motion realism
- Not very accurate
- Perturbation by magnetic fields
- Perturbation introduced by soft tissue and muscle displacement
Optical sensors (infra-red) Exemple:

VICON
Ultrasonic sensors

3 microphones used to identify spatial position of one microphone.

Gyroscopic system

• Sensor composed by 3 gyroscopes along orthogonal axis providing orientation information.
• Gyroscopes are subject to drift so a compass measuring 3D earth magnetic field
Sensing through video-based motion capture (R. Plaenkers, P. Fua)

Much easier in one plane (2D)
Depth cameras: Kinect

- Depth sensor: infrared laser projector combined with monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions
- Kinect: motion sensing input device by Microsoft for the Xbox 360 video game console and Windows PCs.
Sensor Solution

![Kinect](image1)

**RGB Image**

**Depth Image**

**Skeletons**
Body Action Recognition

- Independence of Motion Capture System
- Action recognition solely based on virtual skeleton configuration.
Body Actions

• Action Primitives
  – posture primitives = (position, CoM) or (EEs,position) or (Joints, angles)
  – gesture primitives = (CoM, velocity direction) or (EE, velocity direction)

• ACTION = Boolean Expression of Action Primitives

• e.g. walking = ((spine, forward) AND (left foot, forward)) OR (spine, forward) AND (right foot, forward))
Body Action Recognition
Augmented Reality

Fall Detection

• Robust fall detection system based on marker-less motion capture.
• Novelty: independent of illumination condition, and the person does not require to wear any markers.

Demo: Fall Detection

Fall Detection Based on Markerless Motion Capture
**Face Recognition**

- Automatically identifying a person from a digital image or a video source.

- Typical ways to do this: comparing selected facial features from image and facial database.

- Recognition algorithms: two main approaches:
  - **geometric**, looks at distinguishing features,
  - **photometric**, statistical approach that distils an image into values and compares them with templates to eliminate variances.

- **3D face recognition**; uses 3D sensors to capture information about shape of a face.

- **Skin texture analysis**: captures and uses visual details of skin.
Face Recognition using the Kinect

- **Novelty**: Utilize Local Binary Patterns feature, extremely fast
- **Training**: 20 seconds
- **Recognition**: real-time

Demo: Face Recognition
Body and hand gesture recognition

Hand feature extraction (Cyber-glove)

Upper body joints extraction (Kinect)

Information fusion

Recognize upper body gesture

Yang Xiao; Junsong Yuan; Daniel Thalmann
Human-Virtual Human Interaction by Upper Body Gesture Understanding, Proc. ACM VRST 2013
Upper Body Gestures without Human-object Interaction
Upper Body Gestures with Human-object Interaction
Upper Body Gesture Understanding

\[ F_{\text{hand}} = (h_1, h_2, h_3 \ldots h_{19}, h_{20}) \]

CyberGlove II Data Joints
Upper Body Gesture Understanding

Selected Body Skeletal Joints

\[ J_i = (x_i(t), y_i(t), z_i(t)) \]

\[ J_{ir} = J_i - J_r \]

\[ F_{body} = (J_{1r}, J_{2r}, J_{3r}, J_{4r}) \]
Upper Body Gesture Understanding

Feature Fusion

\[ \hat{F}_i = \frac{F_i - F_i^{\text{min}}}{F_i^{\text{max}} - F_i^{\text{min}}} \]

\[ \hat{F} = (\hat{F}_{\text{hand}}, \hat{F}_{\text{body}}) \]
Human Upper Body Gesture Dataset Samples
Action localization based RGBD cameras

• Problem
  – continuous human-object interaction recognition based on RGBD (Kinect camera)
  – Determine start and end frame of one action

• Motivation
  – Combine skeleton and object information
Orderlet: primitive feature

• Skeleton information

\[ s_i^{(t)} = [x_i^{(t)}, y_i^{(t)}, z_i^{(t)}] \]

- Distance between skeleton point in one frame

\[ \lambda^{(1)} = ||s_i^t - s_j^t||. \]

- Spatial position for each skeleton point in one frame

\[ \lambda^{(2)} = x_i^t \text{ or } y_i^t \text{ or } z_i^t. \]

- Distance between one specific skeleton point in different frames

\[ \lambda^{(3)} = ||s_i^t - s_i^{t-\Delta t}||, \]

• Object Information

- Local Occupancy pattern

\[ \lambda^{(4)} = ||d(i) - d(j)|| = ||l_i - l_j||, \quad 1 \leq i, j \leq N_b, \quad i \neq j. \]
Video Results

Sample Training Sequences of our dataset

Gang Yu, Zicheng Liu, Junsong YuanD, iscriminative Orderlet Mining For Real-time Recognition of Human-Object Interaction, Asian Conference on Computer Vision (ACCV) 2014
Demo: Body gesture recognition
Hand Gesture Recognition
Motivation

- Hand pose estimation has various applications
  - Sign language recognition
  - Virtual environment manipulation
  - Animation synthesis

- Why vision-based solution for bare-hand pose estimation?
  - Solutions with electro-mechanical devices, optical sensors and color gloves are expensive, or inconvenient to use
  - Vision-based method provides cheap and natural human-computer interaction experience

- Challenges in vision-based hand pose estimation
  - Lack of robust and efficient features
  - High degree-of-freedom hand motion
Problem Description

- We focus on vision-based full DOF pose estimation for bare-hand input.
- The hand pose is parameterized as a 27D vector, including the global and local motion. The task is to restore the pose vector for each input frame.

(a) Parameterized pose space
(b) Problem formulation
Hand Gesture Recognition

• Novelty: spatial-temporal feature, which enforces both spatial and temporal constraints in a unified framework for hand parsing and fingertip detection.

• Result: more accurate compared to existing methods.

Hui Liang, et al. 3D Fingertip and Palm Tracking in Depth Image Sequences, ACM International Conference on Multimedia 2012 (MM)
A New Method

- Geodesic distance map
  - The geodesic distance from the palm center to all other points along the hand surface

- Hand parsing
  - Parse the hand region into individual hand parts
  - Combination of temporal and spatial information

- 3D fingertip detection
  - Use of parsing result to guide fingertip detection

- Hand pose estimation
  - Represent the transformed coordinate of the palm by the normal of the palm and the vector pointing from the palm center to the middle fingertip
  - Apply inverse kinematics (IK) to each finger with the fingertip positions to find the finger pose

Demo: Hand Gesture Recognition
Regression forest = set of regression trees

Each tree can retrieve a vote \((v_{ijk}, \rho_{ijk})\) for each pixel \(p_i\) and each joint \(\phi_k\).

New Method

• Task: Predict the sixteen joint locations of the hand from single depth images

• The Regression Forest is utilized to get the per-pixel prediction votes for each hand joint

• **Multimodal Prediction Fusion:** Fuse the per-pixel predictions with the learned hand joint correlations
  – Learning the joint correlations: PCA analysis of the joint parameters $\Phi$ in the training data
  – Seeking the optimal $\Phi^*$ in this reduced dimensional space during testing

Quantitative Results

Training data: 90k synthesized depth images
Testing data: 23k synthesized and 600 real-world depth images
Evaluation Metric: average percentage of the predicted joints within a distance of $D_T$ from the ground truth
Applications
Dolphins for ASD Rehabilitation

- Dolphin-assisted therapy (DAT), a structured program designed for ASD children (USA, Australia)
- Dolphin Encounter – A DAT program in Singapore
- Chia, Kee, Watanabe, & Poh (2009), *Journal of the American Academy of Special Education Professionals*. 
Virtual Pink Dolphins for ASD Rehabilitation

• Immersive Virtual Environment in IMI/NTU
• 3D Virtual Pink Dolphins

• Replace the physical dolphins
• Immersive, Interactive, Serious Games
• Initial research with Special Schools on the use of Virtual Pink Dolphin for ASD Rehabilitation
Crowd Simulation

- Observation of interesting emergent behaviors, e.g., lane formations or panic effects, => crowd motion planning more realistic
Interaction Design

• Natural interface for user
• Device
  – MS Kinect Sensor
• Method
  – Template-based gesture recognition
• Interactions
  – Walk
  – Pick
  – Direct
  – Gather
  – Disperse
  – Lead
  – Stop

Two scenarios

- gathering the agents to a specific orientation.
- making agents disperse after gathering around the avatar
Thank you for your attention..

Questions ?…