Multi-Modal Human-Computer Interaction

Attila Fazekas

University of Debrecen, Hungary
Road Map

- Multi-modal interactions and systems (main categories, examples, benefits)
- Face detection, facial gestures recognition, etc.
- Body gesture detection
- Examples
• There are two views on multi-modal interaction:
  • The first focuses on the human side: perception and control. There the word modality refers to human input and output channels.
  • The second view focuses on synergistic using two or more computer input or output modalities.
The Modalities

- We can divide the modalities in seven groups
- Internal chemical (blood oxygen, etc.)
- External chemical (taste, etc.)
- Somatic senses (touch, etc.)
- Muscle sense (stretch, etc.)
- Sense of balance
- Hearing
- Vision
Definition of the Multimodality

- "Multi-modality is the capacity of the system to communicate with a user along different types of communication channels."

- Both multimedia and multi-modal systems use multiple communication channels. But a multi-modal system strives for meaning.
Two Types of Multi-modal Systems

- The goal is to use the computer as a tool.
- The computer as a dialogue partner.
• Bolt's Put-That-There system. In this system the user could move objects on screen by pointing and speaking.

• CUBRICON is a system that uses mouse pointing and speech.

• Oviatt presented a multi-modal system for dynamic interactive maps.
Benefits

- **Efficiency** follows from using each modality for the task that it is best suited for.
- **Redundancy** increases the likelihood that communication proceeds smoothly because there are many simultaneous references to the same issue.
- **Perceptability** increases when the tasks are facilitated in spatial context.
Benefits

• **Naturalness** follows from the free choice of modalities and may result in a human-computer communication that is close to human-human communication.

• **Accuracy** increases when another modality can indicate an object more accurately than the main modality.
Applications

- Mobile telecommunication
- Hands-free devices to computers
- Using in a car
- Interactive information panel
Face Analysis

- Face carries a lot of important information in communication
- Monitoring the face is fundamental in HCI
- First step: face detection (localization)
- Using the localized face can be performed:
  - Tracking face and facial features
    - 2D face tracking, gaze estimation, head-shake detection
  - Face classification
    - gender, age, facial expressions, race
  - Feature extraction
    - skin/eye/hair color
    - mustache/beard detection
Face Detection/Tracking

- Viola and Jones detector is used
  - Face detection is reduced to image classification problem

- Given a set of feature types:

- Training:
  - positive (faces) and negative (random images) examples
  - those features are selected which fits on the positive set (finding position and their extent)
  - the selected features are collected into a cascade file

- Face detection
  - the different scale of the input image is scanned through
  - fitting the set of features:
    - if all are fitting => there is a face
Face Classification

- Can be used for: gender, age, facial expression, race detection
- Preprocessing - for feature extraction
  - Cutting the face
  - LBP transform
  - Gabor transform
- Classification (using the model file)
  - SVM
  - AdaBoost
- Databases: Cohn-Kanade, FERET
- Currently we have the following classifiers trained (LBP+SVM)
  - Gender (male vs. female)
  - Facial expressions (happy, sad, surprised, angry)
  - Age estimation (10-29, 30-49, 50-)
  - Race (asian, hispanic, black, white)
Facial Feature Color Extraction

- Determining the color of various facial features: skin, hair and eyes.
  - The full color range of the segmented face image will be reduced to color categories based on human cognition principles.
  - The segmentation steps of the HI plane in case of hair colors: (a) color marker points, (b) convex hull, (c) distance transform and (d) the segmented plane:
• The colors of the facial features are determined in two steps:
  - First, the skin, eyes and hair are segmented in the image using only structural information
  - Then, within the segmented regions the huge number of colors in real color images is substituted by a smaller color set, which is used to determine the color of a given feature.
Head-pose Tracking, Gaze Estimation

- **Head pose estimation**
  - Detecting the location of the facial features using individual feature detectors (Viola and Jones detectors):
    - eyes, mouth, nose tip
  - Based on the position of the facial features, the POSIT algorithm is used to estimate the three rotation and the translation vector of the head pose

- **Further work:**
  - Pupil tracking
  - Combining head pose with the position of the pupil for gaze estimation.
Body Gestures on Stereo Image

- **Goal:** estimate depth based on an image pair
- **Parallel stereo configuration**

\[ \text{Disparity: } d = x_l - x_r \]

\[ \text{Depth estimation: } D \approx \frac{\lambda bf}{d} \]

- \( C_l, C_r \) - cams
- \( f \) - focal length
- \( b \) - baseline
- \( \lambda \) - ratio: pixels/metre
- \( p_l(x_l, y_l) \)
- \( p_r(x_r, y_r) \)
Computing Disparity

- Feature based approach
  - detect feature points (corners) on the both images
  - compute correlation between feature points on the different image
  - define stereo pairs
  - extend values (region growing, interpolation)
Computing Disparity

Problems:
- Too few matching pairs are there on the torso.
- Wrong matching on the similar areas.
- Decrease compute time:
  - Compute disparity on the important part (ellipse shape)
  - Computation disparity in each intersection of 4. row and 4. column within ellipse, and extend disparity.
  - time: ~35ms
• **Window based approach**
  - Compute correlation between all the points pair of images at the same line.
  - Common correlation measure is Sum of Squared Difference

\[
SSD(u, v) = \sum_{x, y \in W(u, v)} [L(x, y) - R(x - d, y)]^2
\]

- SSD is sensitiv to lighting condition
- Our measure:

\[
S(u, v) = \sum_{x, y \in W(u, v)} |(L(x, y) - L(u, v)) - (R(x - d, y) - R(u - d, v))|
\]
Computing Disparity

- Quality is depends on the size of window.
- Test:
  - Image size: 320x240
  - disparity range: 15-45
  - Window size: 11x11
  - time: ~900ms
Chess Player Turk-2

SSIP’11
08.07.2011
Thank you for your attention!