CS 160: Lecture 23

Professor John Canny
Spring 2003
Preamble

Handout for next lecture.

Quiz on Help systems.
Multimodal Interfaces

- Multi-modal refers to interfaces that support non-GUI interaction.

- Speech and pen input are two common examples - and are complementary.
Speech+pen Interfaces

- Speech is the preferred medium for subject, verb, object expression.

- Writing or gesture provide locative information (pointing etc).
Speech+pen Interfaces

Speech+pen for visual-spatial tasks (compared to speech only)
* 10% faster.
* 36% fewer task-critical errors.
* Shorter and simpler linguistic constructions.
* 90-100% user preference to interact this way.
Put-That-There

User points at object, and says “put that” (grab), then points to destination and says “there” (drop).

* Very good for deictic actions, (speak and point), but these are only 20% of actions. For the rest, need complex gestures.
Multimodal advantages

Advantages for error recovery:
* Users intuitively pick the mode that is less error-prone.
* Language is often simplified.
* Users intuitively switch modes after an error, so the same problem is not repeated.
Multimodal advantages

Other situations where mode choice helps:
* Users with disability.
* People with a strong accent or a cold.
* People with RSI.
* Young children or non-literate users.
Multimodal advantages

For collaborative work, multimodal interfaces can communicate a lot more than text:

* Speech contains prosodic information.
* Gesture communicates emotion.
* Writing has several expressive dimensions.
Multimodal challenges

Using multimodal input generally requires advanced recognition methods:
* For each mode.
* For combining redundant information.
* For combining non-redundant information: “open this file (pointing)”

Information is combined at two levels:
* Feature level (early fusion).
* Semantic level (late fusion).
Early fusion

- Vision data
- Speech data
- Other sensor data

Feature recognizer

Action recognizer

Fusion data
Early fusion

Early fusion applies to combinations like speech+lip movement. It is difficult because:

* Of the need for MM training data.
* Because data need to be closely synchronized.
* Computational and training costs.
Late fusion

Vision data

Feature recognizer

Action recognizer

Speech data

Feature recognizer

Action recognizer

Other sensor data

Feature recognizer

Action recognizer

Fusion data

Recognized Actions
Late fusion

Late fusion is appropriate for combinations of complementary information, like pen+speech.

- Recognizers are trained and used separately.
- Unimodal recognizers are available off-the-shelf.
- It's still important to accurately time-stamp all inputs: typical delays are known between e.g. gesture and speech.
Contrast between MM and GUIs

- GUI interfaces often restrict input to single non-overlapping events, while MM interfaces handle all inputs at once.

- GUI events are unambiguous, MM inputs are based on recognition and require a probabilistic approach.

- MM interfaces are often distributed on a network.
Agent architectures

- Allow parts of an MM system to be written separately, in the most appropriate language, and integrated easily.

- OAA: Open-Agent Architecture (Cohen et al) supports MM interfaces.

- Blackboards and message queues are often used to simplify inter-agent communication.
  * Jini, Javaspaces, Tspaces, JXTA, JMS, MSMQ...
Final project presentations on May 12 and 13.

Presentations go by group number. Groups 7-12 on Monday 12, groups 1-6 on Tuesday 13.

Final reports are due on Weds May 7.
Symbolic/statistical approaches

- Allow symbolic operations like unification (binding of terms like “this”) + probabilistic reasoning (possible interpretations of “this”).

- The MTC system is an example
  - Members are recognizers.
  - Teams cluster data from recognizers.
  - The committee weights results from various teams.
MTC architecture
Probabilistic Toolkits

- The “graphical models toolkit” U. Washington (Bilmes and Zweig).
  * Good for speech and time-series data.

- MSBNx Bayes Net toolkit from Microsoft (Kadie et al.)

- UCLA MUSE: middleware for sensor fusion (also using Bayes nets).
MM systems

Designers Outpost (Berkeley)
MM systems: Quickset (OGI)

- Other QuickSet Users
- ModSAF Simulations
- 3-D "cave" interaction
- Spoken Language
- Handheld Computers
- Gesture
Crossweaver (Berkeley)
Crossweaver (Berkeley)

- Crossweaver is a prototyping system for multi-modal (primarily pen and speech) UIs.

- Also allows cross-platform development (for PDAs, Tablet-PCs, desktops.)
Summary

- Multi-modal systems provide several advantages.
- Speech and pointing are complementary.
- Challenges for multi-modal.
- Early vs. late fusion.
- MM architectures, fusion approaches.
- Examples of MM systems.