Perceptual Interfaces

Adapted from
Matthew Turk's (UCSB) and
George G. Robertson's (Microsoft Research)
slides on perceptual interfaces

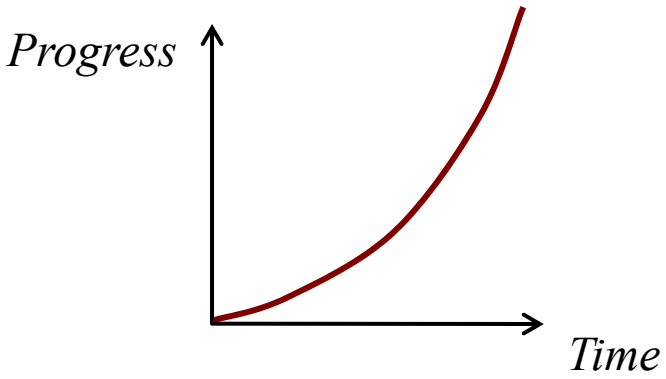
Outline

- ✓ Why Perceptual Interfaces?
- ✓ Multimodal interfaces
- ✓ Vision Based Interfaces (VBI)
- ✓ Examples

Observation

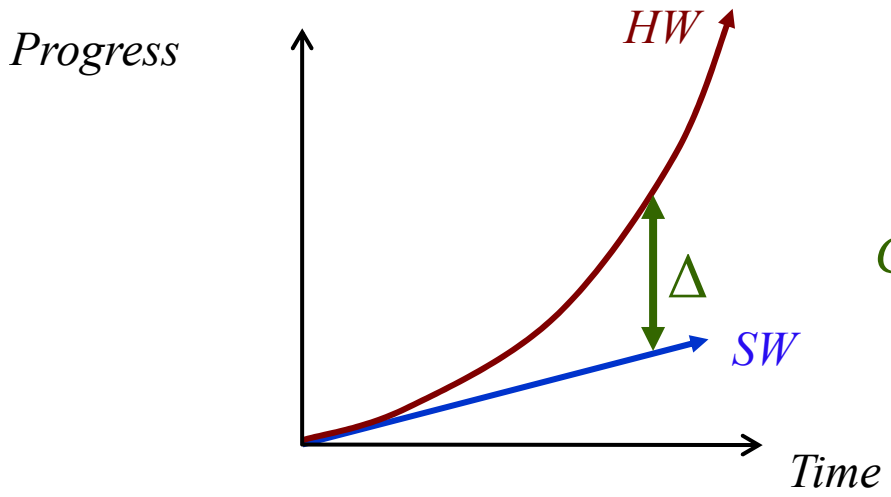
- **Moore's Law** has driven computer technology for decades

  Exponential improvement in HW
  - 5 years ~ 10x improvement
  - 10 years ~ 100x improvement
  - 20 years ~ 10,000x improvement



- But… there has been no Moore's Law for user interfaces!
  - The result?

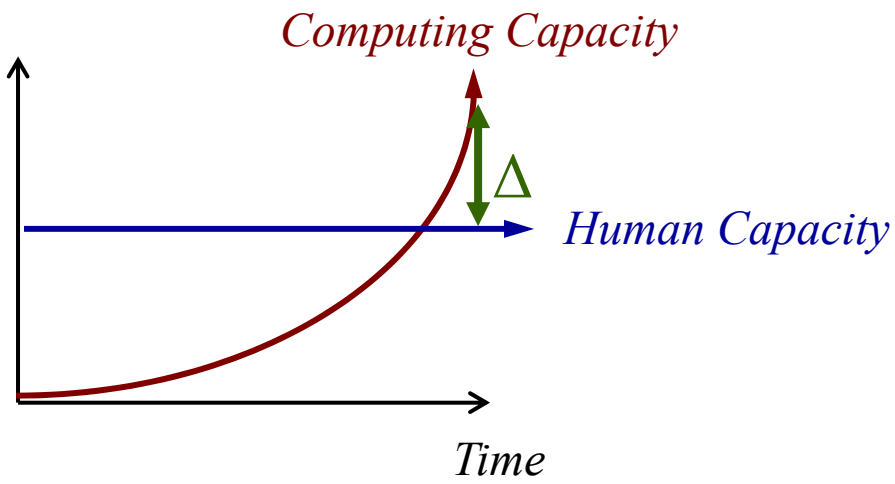
The result



Curse of the delta!

Another view:

There's no Moore's Law for people!



Curse of the delta

Evolution of user interfaces

When	Implementation	Paradigm
1950s	Switches, punched cards	None
1970s	Command-line interface	Typewriter
1980s	Graphical UI (GUI)	Desktop
2000s	???	???

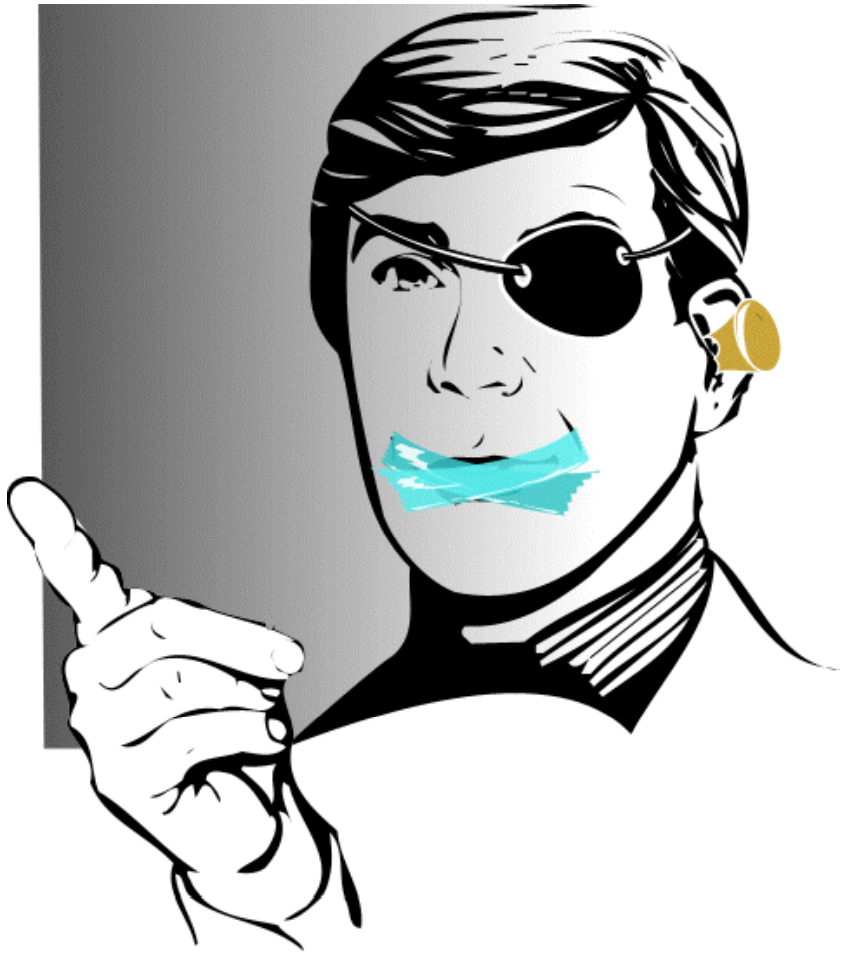
Current UI Limitations

Failure to use Human Abilities

Limited Vision (Flat, 2D)

No Speech

No Gestures



Limited Audio

One Hand Tied Behind Back

Limited Tactile

The Next Big Thing in UI?

- Immersive environments
  - Wearable computers, Virtual Reality, Augmented Reality…
- Ubiquitous Computing
  - Invisible, pervasive
- Tangible UI
  - Coupling of physical objects and digital data
- Multimodal UI
  - Sound, speech, gesture…
- Affective Computing
  - Computers that understand and express emotion

Evolution of user interfaces

When	Implementation	Paradigm
1950s	Switches, punched cards	None
1970s	Command-line interface	Typewriter
1980s	Graphical UI (GUI)	Desktop
2000s	Perceptual UI (PUI)	Natural interaction

Perceptual Interfaces

Highly interactive, multimodal interfaces modeled after natural human-to-human interaction

- Goal: For people to be able to interact with computers in a similar fashion to how they interact with each other and with the physical world

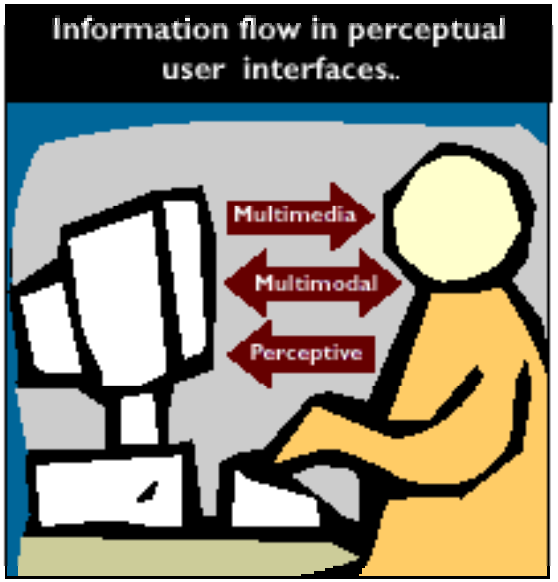
Not just passive

Multiple modalities, not just mouse, keyboard, monitor

“Perceptual” User Interfaces

- **Perceptive**
  - human-like perceptual capabilities (what is the user saying, who is the user, where is the user, what is he doing?)
- **Multimodal**
  - People use multiple modalities to communicate (speech, gestures, facial expressions, …)

- Multimedia
  - Text, graphics, audio and video



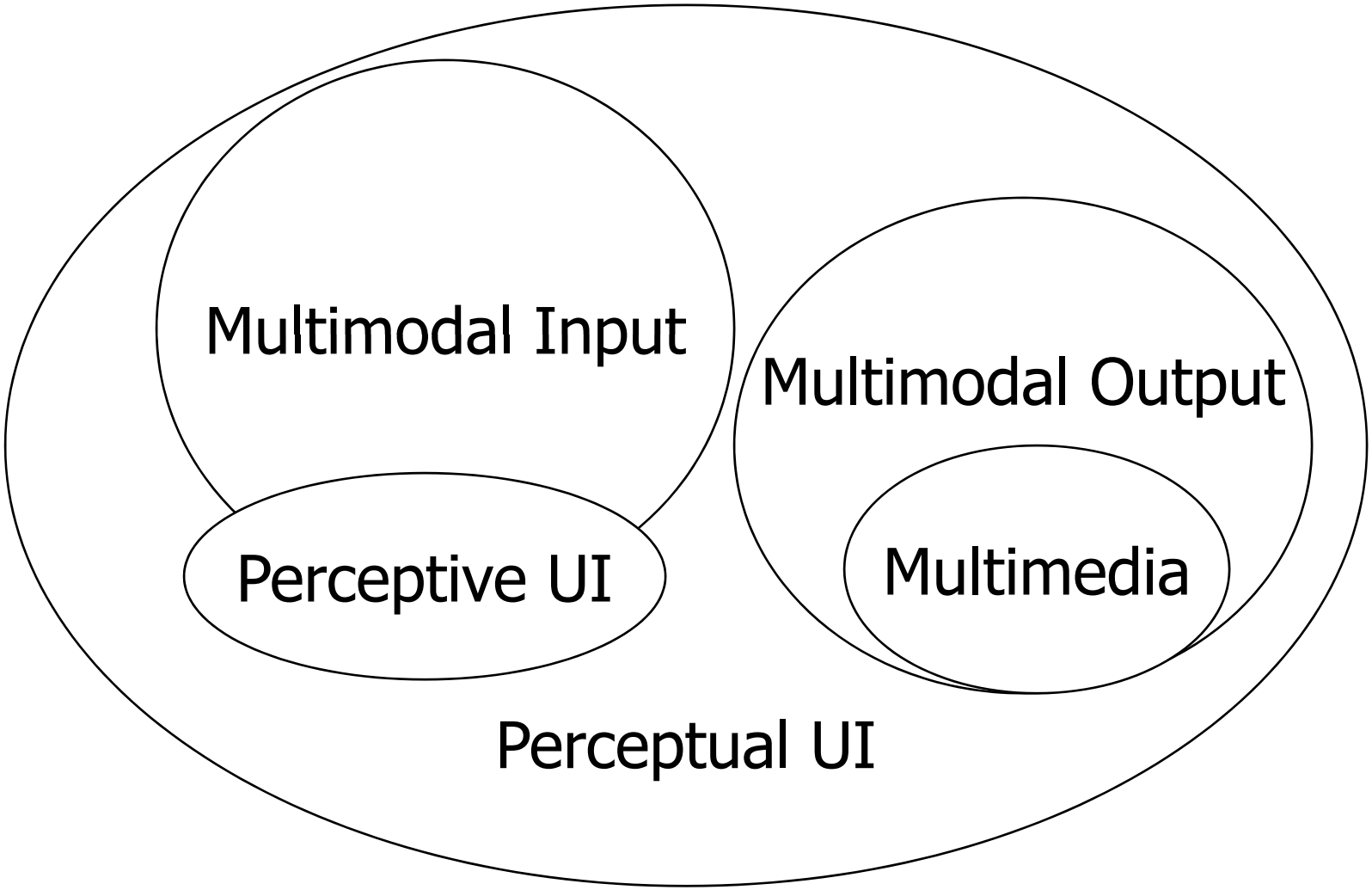
Perception

- In order to respond appropriately, objects/room need(s) to pay attention to
  - **People** and
  - **Context**

- Machines have to be *aware* of their environment:
  - **Who, What, When, Where and Why**?

- Interfaces must be **adaptive** to
  - Overall situation
  - Individual User

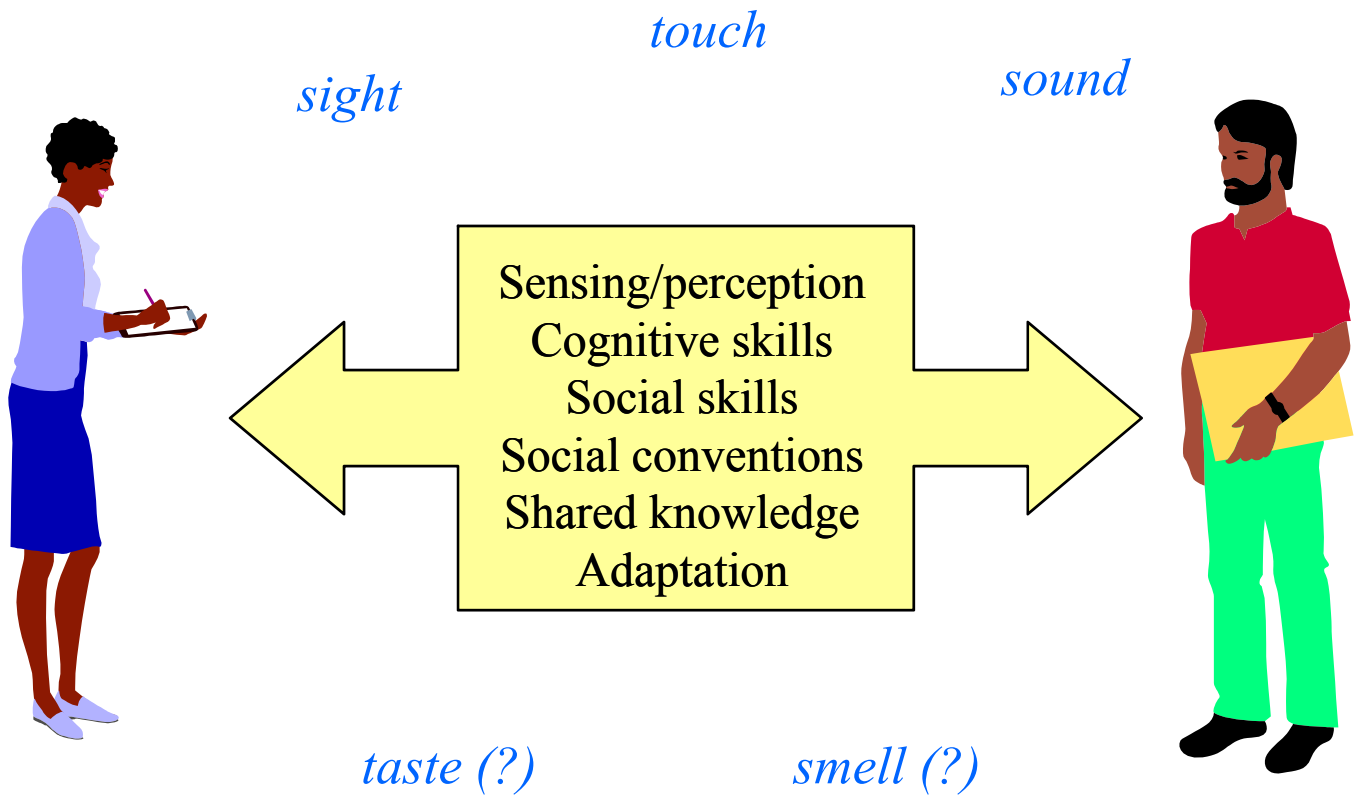
How Do The Pieces Fit?



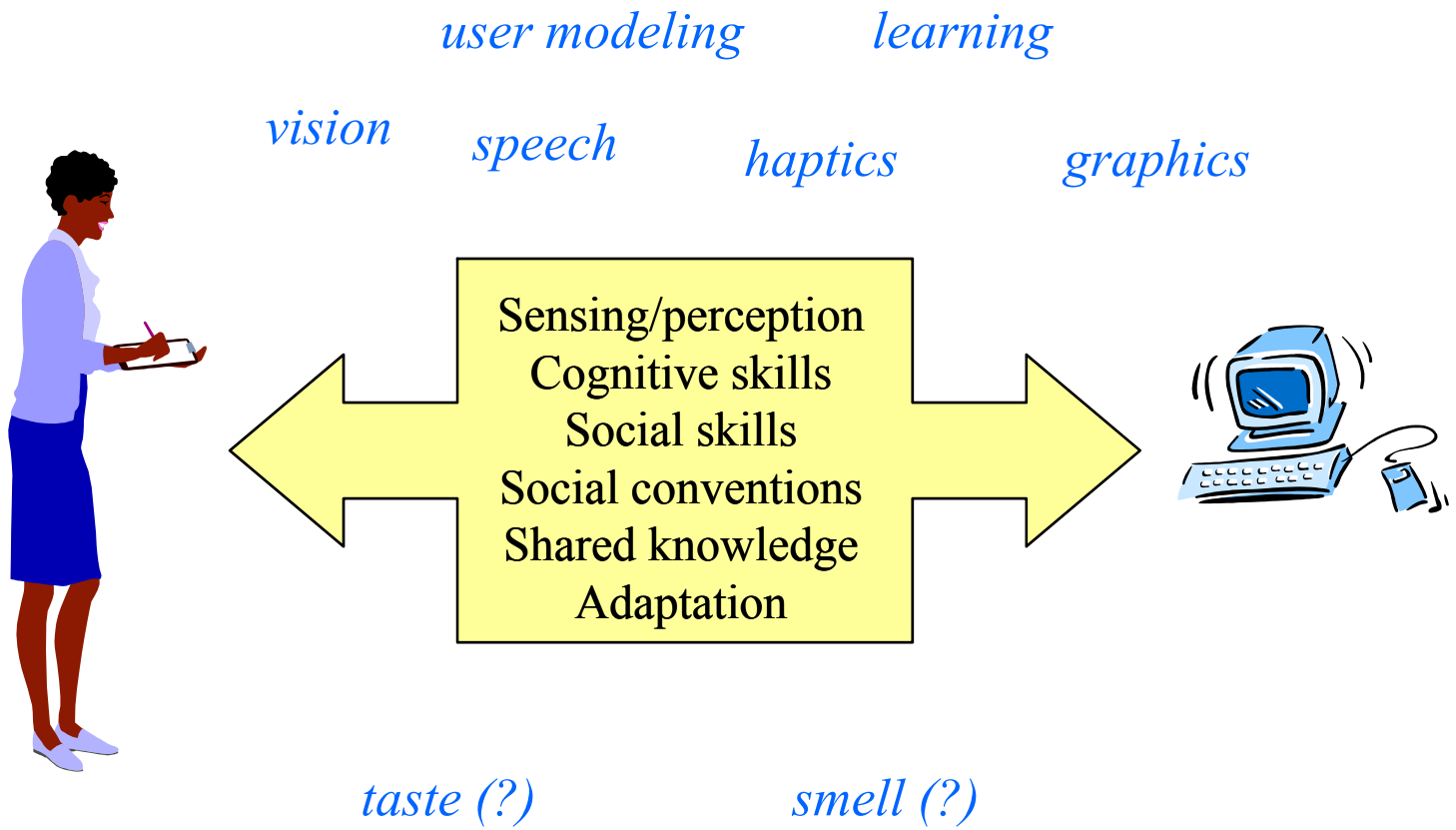
Perceptual User Interfaces (PUI)

- Special section on PUIs in the March 2000 issues of *Communications of the ACM*, edited by Matthew Turk and George Robertson.
- PUIs combine natural human capabilities of communication, motor, cognitive, and perceptual skills with computer I/O devices, machine perception, and reasoning.
- Integrate research results from different disciplines
  - vision, speech, graphics and visualization, user modeling, haptics, and cognitive psychology

Natural human interaction



Perceptual Interface

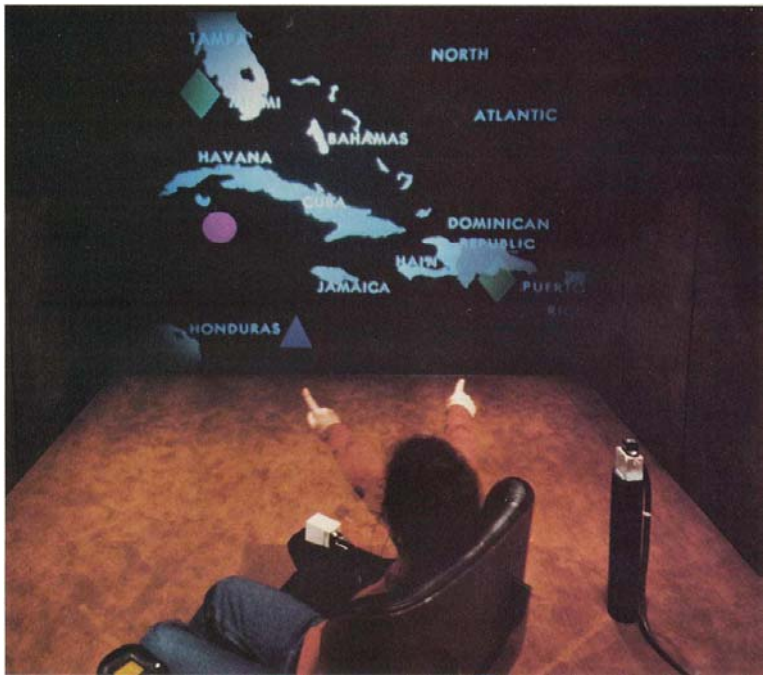


What are Multimodal Interfaces?

- Attempts to use human communication skills
- Provide user with multiple modalities
- May be simultaneous or not
- Fusion vs. Temporal Constraints
- Multiple styles of interaction

Early example

"Put That There" (Bolt 1980)…



Speech and gestures used simultaneously

Why Multimodal Interfaces?

- Today's interfaces fall far short of human capabilities
  - Higher bandwidth is possible
  - Different modalities excel at different tasks
  - Errors and disfluencies reduced
- Multimodal interfaces are more engaging
  - Users perceived multiple things at once
  - User do multiple things at once

Motivation: Why PUIs?

- Many reasons, including:
  - The "glorified typewriter" GUI model is too weak, too constraining, for the ways we will use computers in the future
  - One size doesn't fit all – diverse HCI requirements from small mobile devices to larger powerful embedded devices.
  - Transfer of natural, social skills – easy to learn
  - Simplicity: simple = natural, adaptive
  - Technology is coming: no longer deaf, dumb, and blind
  - To enable both *control* and *awareness*

How could we do this?

- Develop and integrate various relevant technologies, such as:

Speech recognition

Speech synthesis

Natural language processing

Vision (recognition and tracking)

Graphics, animation, visualization

Haptic I/O

Affective computing

Tangible interfaces

Sound recognition

Sound generation

User modeling

Conversational interfaces

Detecting gesture

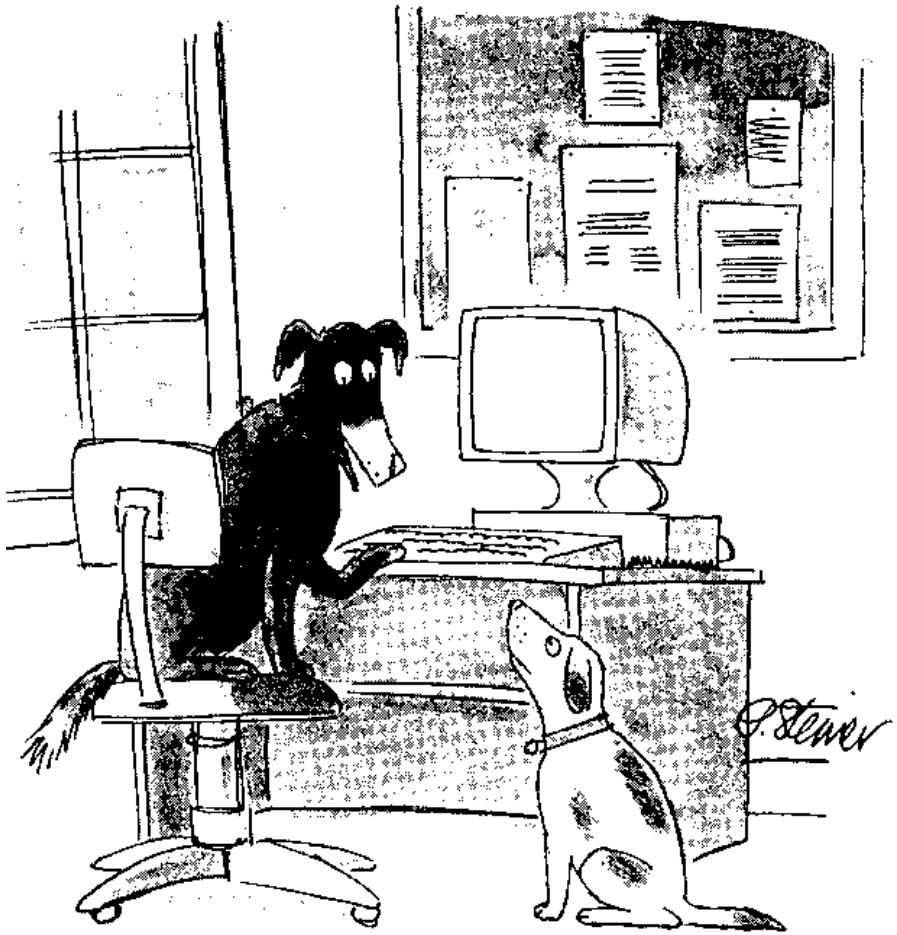
Being aware of the user

Natural navigation

There are many issues!

- What are the appropriate and most useful input/output modalities? (vision, speech, haptic, *taste, smell?)*
- Is the event-based model appropriate?
- What is a perceptual event?
- Is there a useful, reliable subset?
- Non-deterministic events
- Future progress (expanding the event set)
- Allocation of resources
- Multiple goal management
- Training, calibration
- Quality and control of sensors
- Environment restrictions
- Privacy

Issues (cont.)



"On the Internet, nobody knows you're a dog."

New Yorker, 5-Jul-1993, p. 61

Some PUI objections

- Arguments against intelligent, adaptive, agent-based, and anthropomorphic interfaces

- HCI should be characterized by:
  - Direct manipulation
  - Predictable interactions
  - Giving responsibility and a sense of accomplishment to users

- Won't work – "AI hard"
  - Is 50% of HAL good enough?

Two major obstacles

- Technology (the easy one)
  - Lots of researchers worldwide
  - Increasing interest
  - Consistent progress

- The Marketplace (the hard one)
  - But there's growing convergence: hw/sw advances, commercial interest in biometrics, accessibility, recognition technologies, virtual reality, entertainment….

but still... not quite there yet...

versus

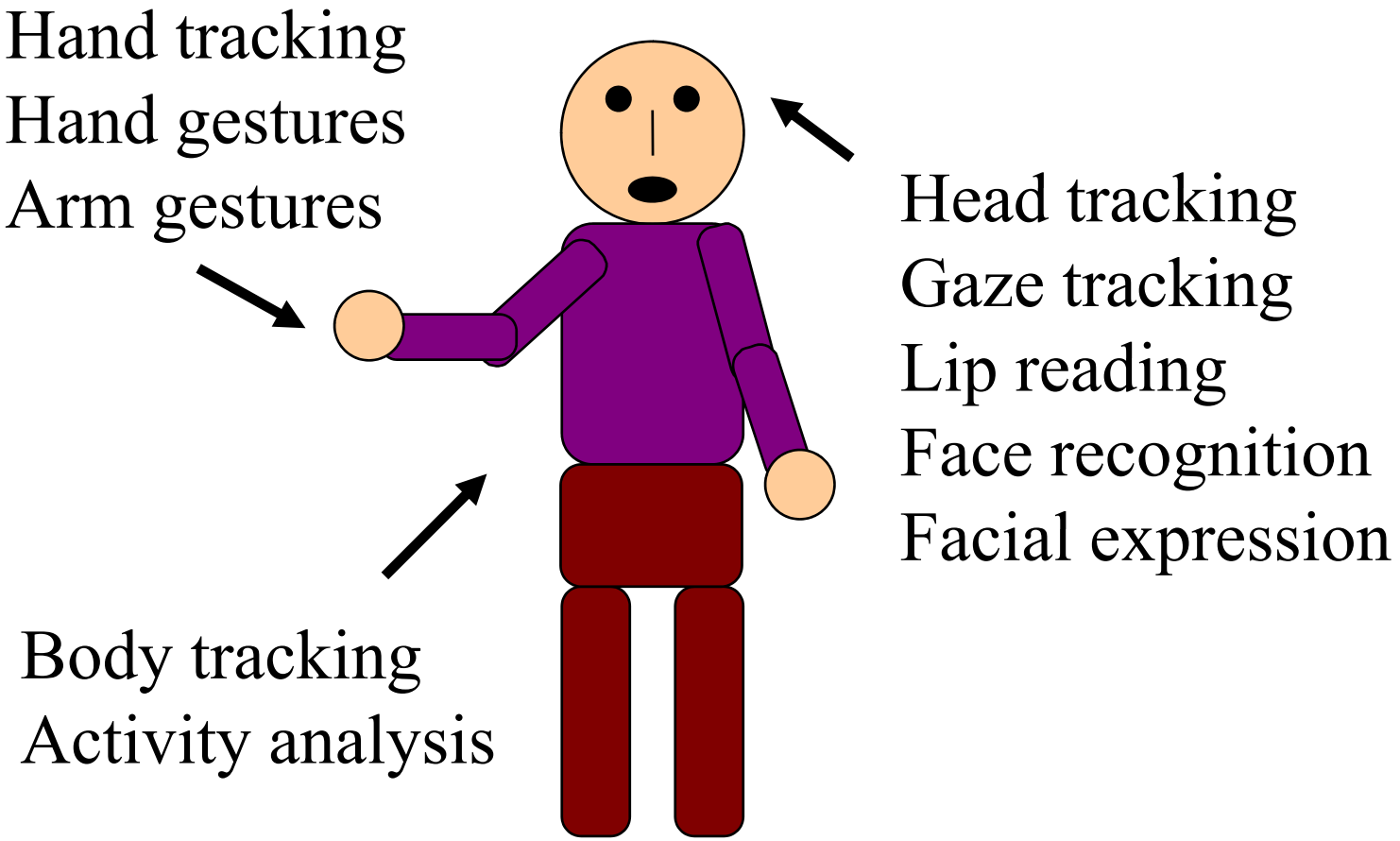
Figure 4. The author wearing a variety of new devices. The glasses (built by Microoptical, Boston) contain a computer display nearly invisible to others. The jacket has a keyboard literally embroidered into the cloth. The lapel has a context sensor that classifies the user's surroundings. And, of course, there's a computer (not visible in this photograph).

Vision Based Interfaces (VBI)

- Visual cues are important in communication!
- Useful visual cues
  - Presence
  - Location
  - Identity (and age, sex, nationality, etc.)
  - Facial expression
  - Body language
  - Attention (gaze direction)
  - Gestures for control and communication
  - Lip movement
  - Activity

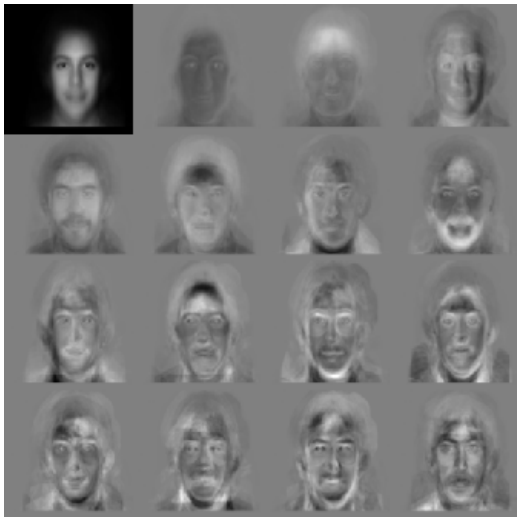
VBI – using computer vision to perceive these cues

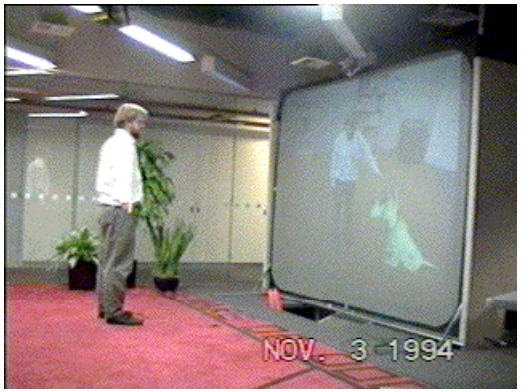
Elements of VBI

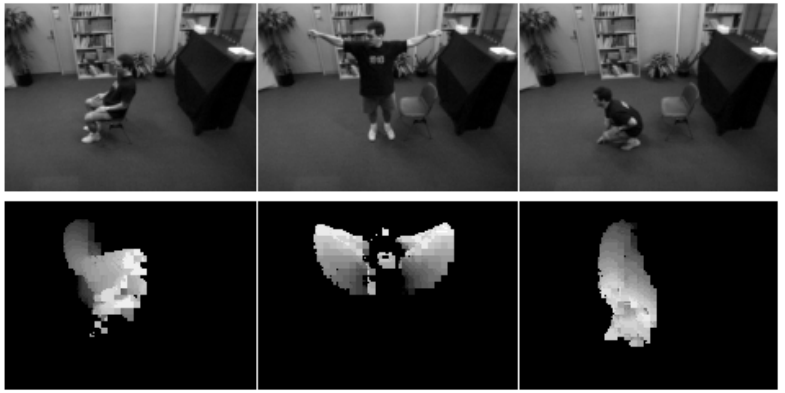


Some VBI application areas

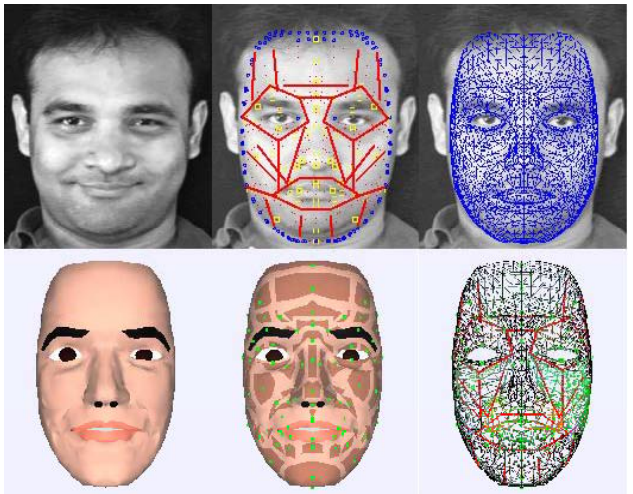
- Accessibility, hands-free computing
- Game input
- Social interfaces
- Teleconferencing
- Improved speech recognition (speechreading)
- User-aware applications
- Intelligent environments
- Biometrics
- Movement analysis (medicine, sports)







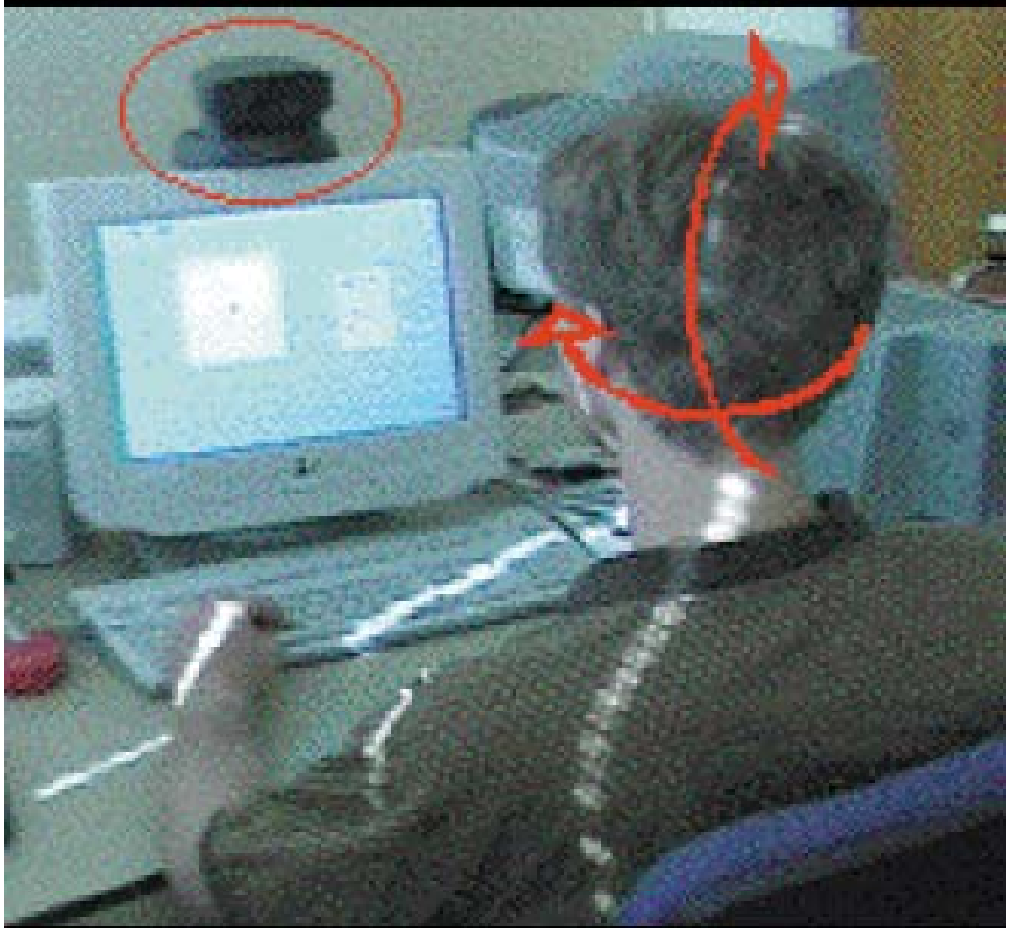
MIT Media Lab
1990s



Perceptual Window

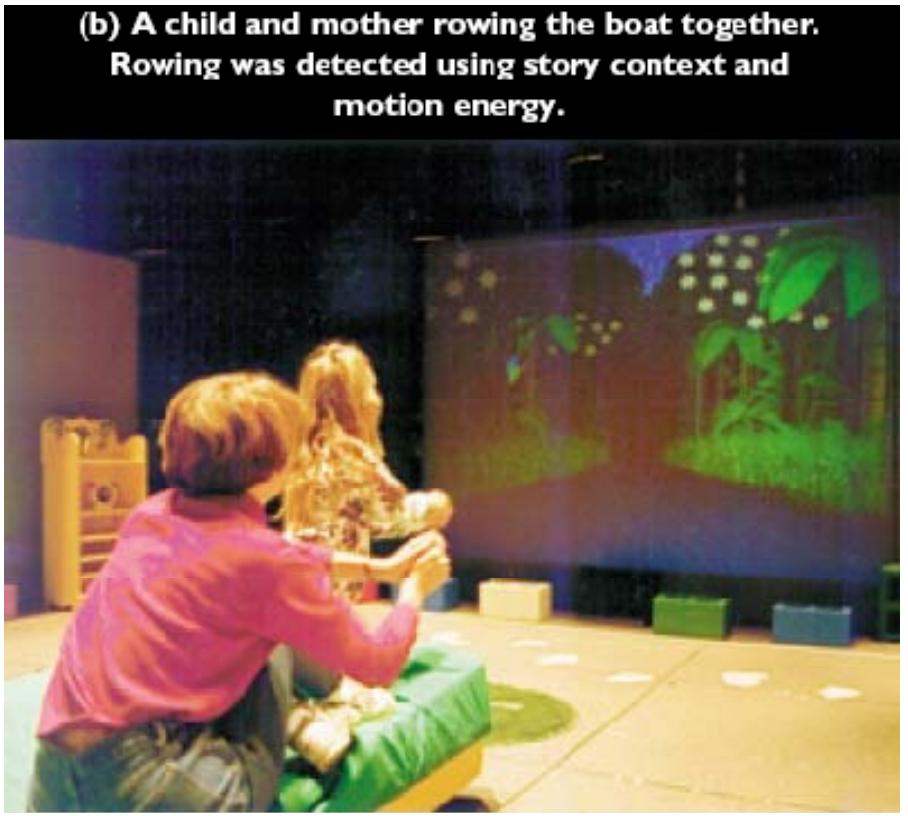
- Hand and mouse form the dominant stream
- Head is used as non-dominant stream
- Better than eye tracking
  - Fixation and saccades

Figure 2. The Perceptual Window uses small head motions as a second input stream to navigate within a document.



KidsRoom (Bobick et al 2000)

(a) A view of the KidsRoom showing the two projection screens and the movable bed.



(b) A child and mother rowing the boat together. Rowing was detected using story context and motion energy.

The technology

- Tracking faces
  - tracking the whole face, lips, gaze, or focus of attention
- Tracking bodies
  - person tracking
- Combining audio info with lip tracking info

Tracking of Human Faces

- A face provides different functions:
  - identification
  - perception of emotional expressions

- Tracking of faces:
  - lip-reading
  - eye/gaze tracking
  - facial action analysis / synthesis

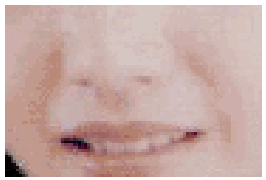
Color Based Face Tracking

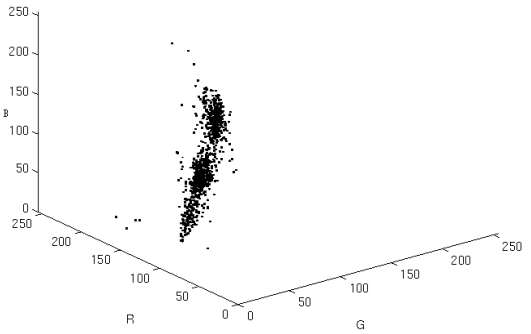
Human skin-colors:

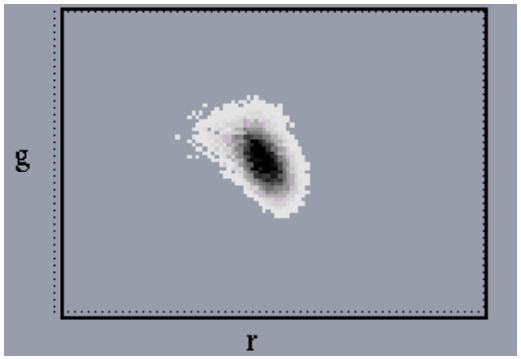
- cluster in a small area of a color space
- skin-colors of different people mainly differ in intensity!
- variance can be reduced by color normalization
- distribution can be characterized by a Gaussian model

Chromatic colors:

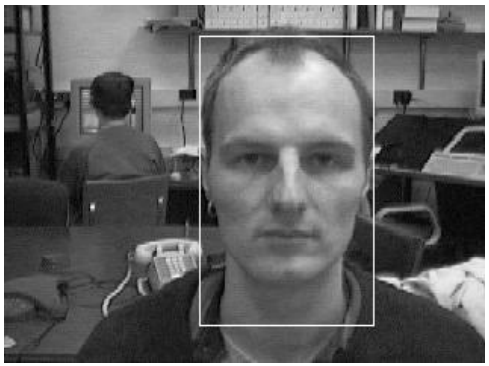
$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B}$$

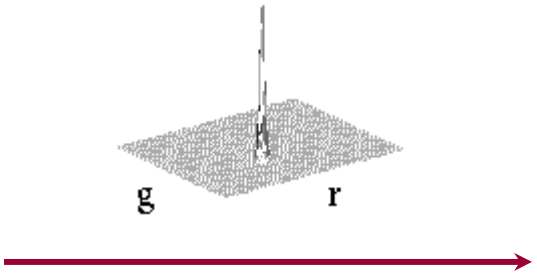






Color Model





Advantages:

- very fast
- orientation invariant
- stable object representation
- not person-dependent
- model parameters can be quickly adapted

Disadvantages:

- environment dependent
- (light-sources heavily affect color distribution)

Tracking Gaze and Focus of Attention

- In meetings:
  - to determine the addressee of a speech act
  - to track the participants attention
  - to analyze, who was in the center of focus
  - for meeting indexing / retrieval
- Interactive rooms
  - to guide the environments focus to the right application
  - to suppress unwanted responses
- Virtual collaborative workspaces (CSCW)
- Human-Robot Cooperation
- Cars (Driver monitoring)

Head Pose Estimation

- Model-based approaches:
  - Locate and track a number of facial features
  - Compute head pose from 2D to 3D correspondences (Gee & Cipolla '94, Stiefelhagen et.al '96, Jebara & Pentland '97,Toyama '98)

- Example-based approaches:
  - estimate new pose with function approximator
  - use face database to encode images (Pentland et.al. '94)

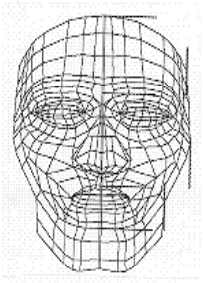
Model-based Head Pose estimation

•Find correspondences between points in a 3D model and points in the image

• Iteratively solve linear equation system to find pose parameters (r_x, r_y, r_z, t_x, t_y, t_z)

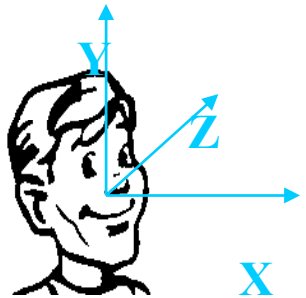
Image

Feature Tracking



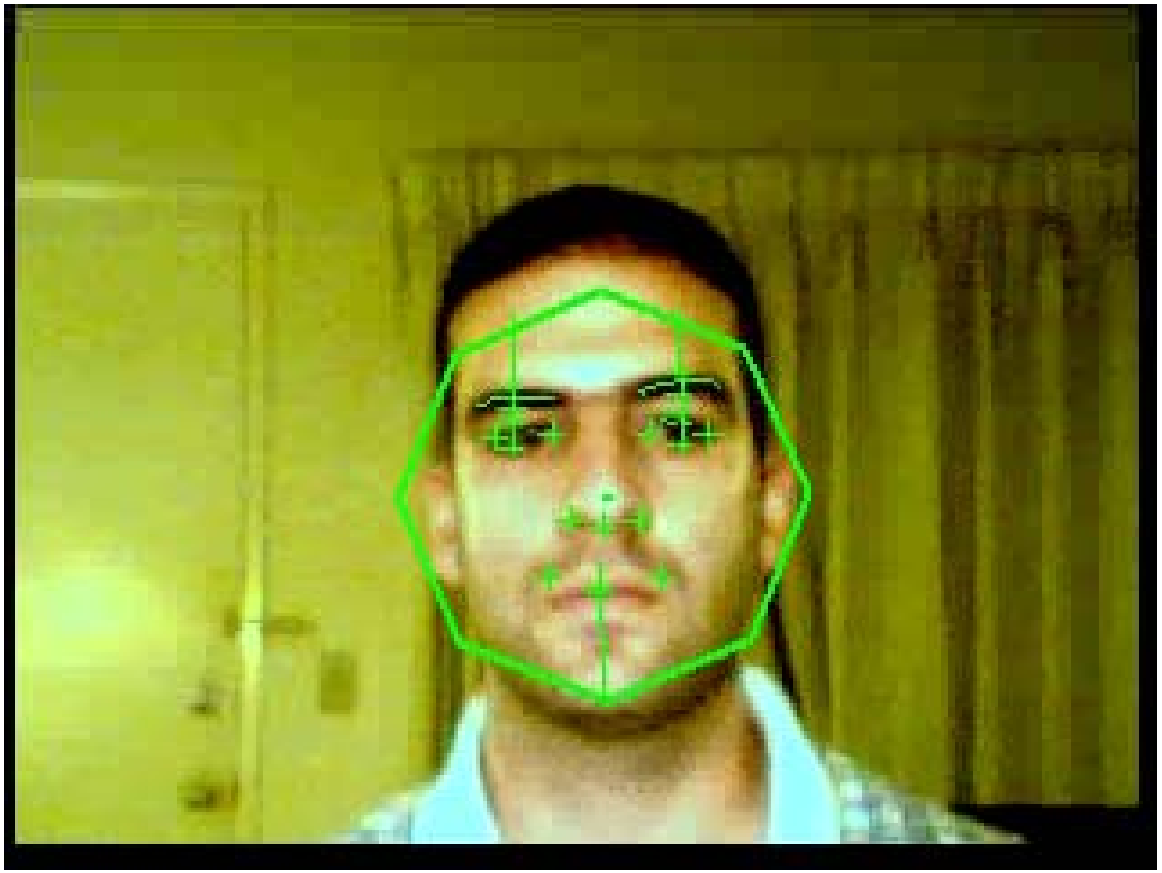
3D Model

Pose Estimation



Real World

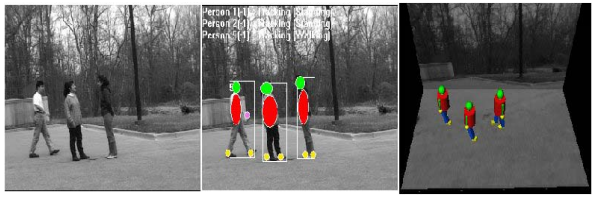
Head tracking demo

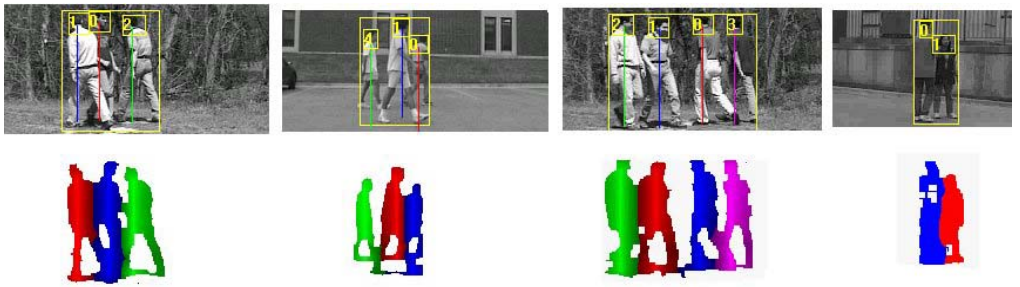


Person Tracking

Vision based localization of people/objects:

- Single Perspective:
- Multiple Perspective:





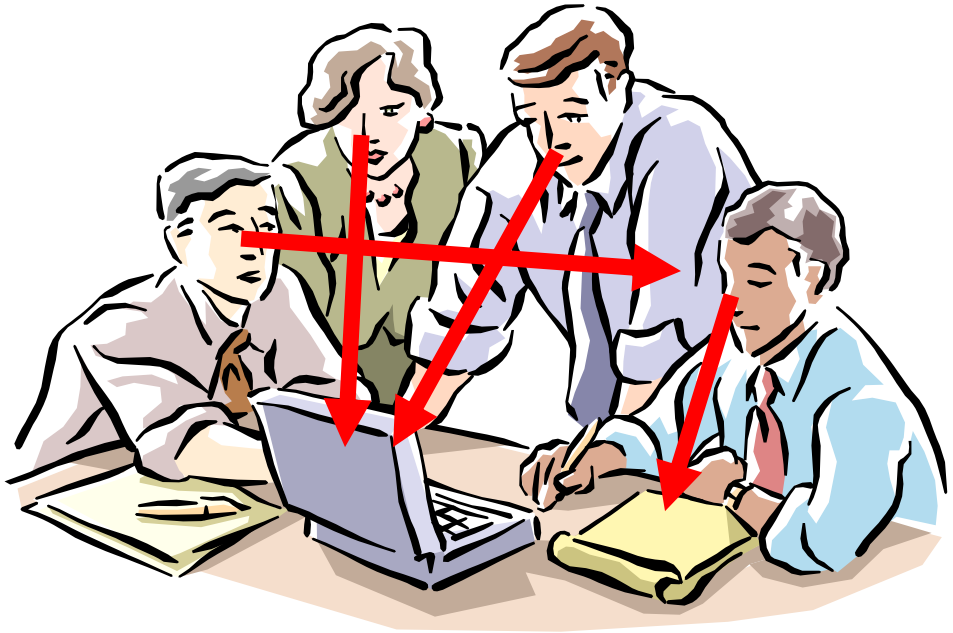
More examples

- Some applications from UCSB Four Eyes lab
- 4 I's: **I**maging, **I**nteraction, and **I**nnovative **I**nterfaces

- Research in computer vision and human-computer interaction
  - Vision based and multimodal interfaces
  - Augmented reality and virtual environments
  - Multimodal biometrics
  - Wearable and mobile computing
  - 3D graphics
  - ….

1. Coarse face direction

- Problem: Coarsely track multiple, possibly low-resolution face images in a scene
- Goal: Capture group behavior (attention); real-time
  - Estimate the "Focus of Intention" (attention + semantics)

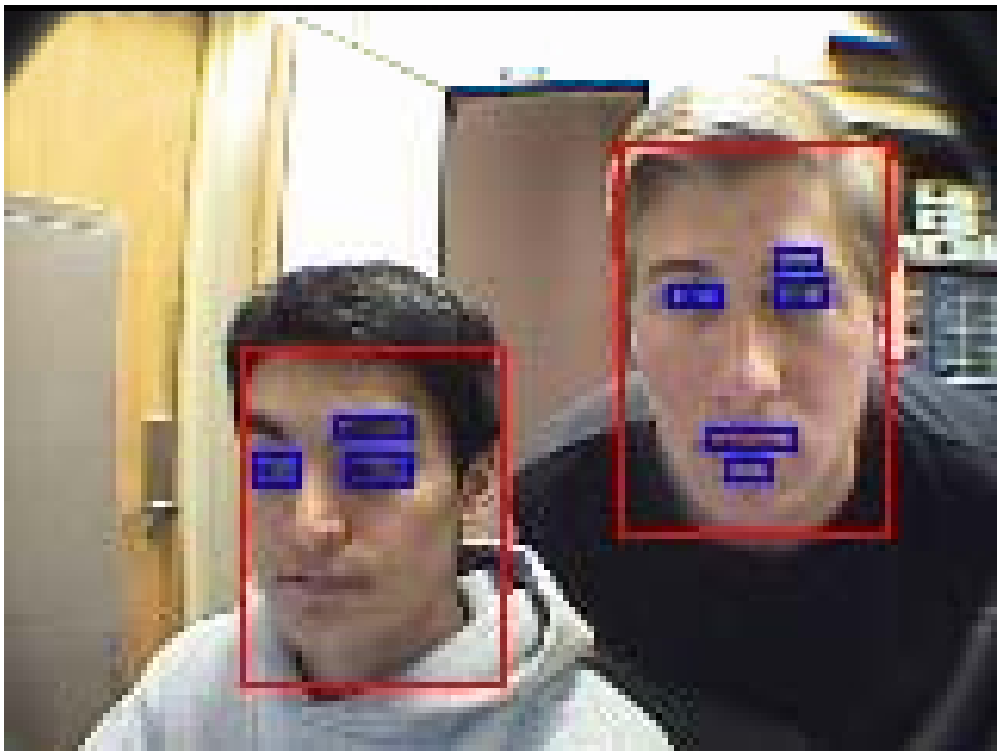


Action understanding
Meeting annotation
Audience feedback
Videoconferencing
Etc.

Coarse face direction (cont.)

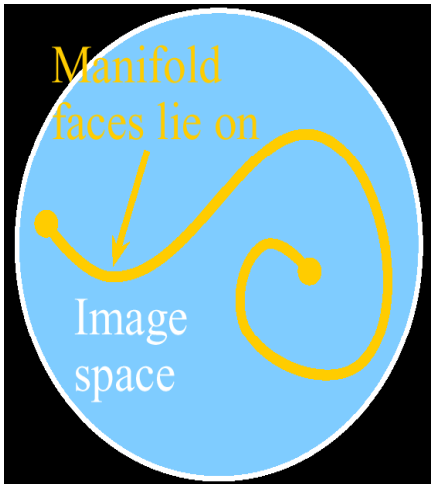
- Strategy:
  - Fast color-based skin tracking
  - Simple feature location
    - Non-skin areas
  - Simple statistics
  - Look for correlation with head direction (relative to camera)
  - f (*statistical measures*) = *direction*

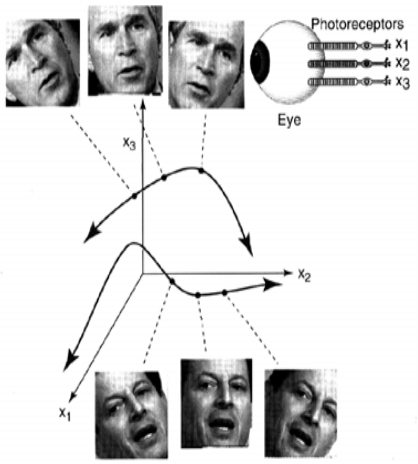
Example results



2. Facial expression analysis

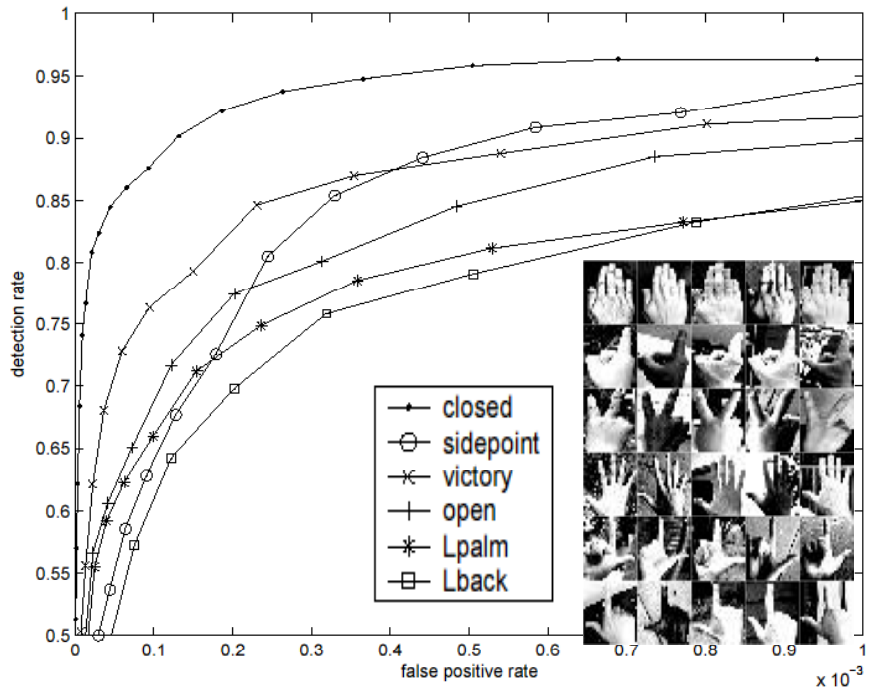
- Facial expression representation and visualization
- Use non-linear manifolds to represent dynamic facial expressions
- Intuition:
  - The images of all facial expressions by a person makes a smooth manifold in (high-dimensional) image space, with the "neutral" face as the central reference point.

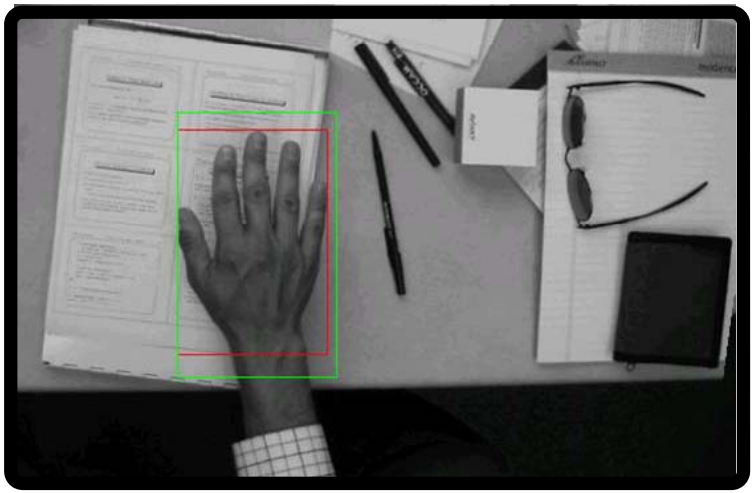


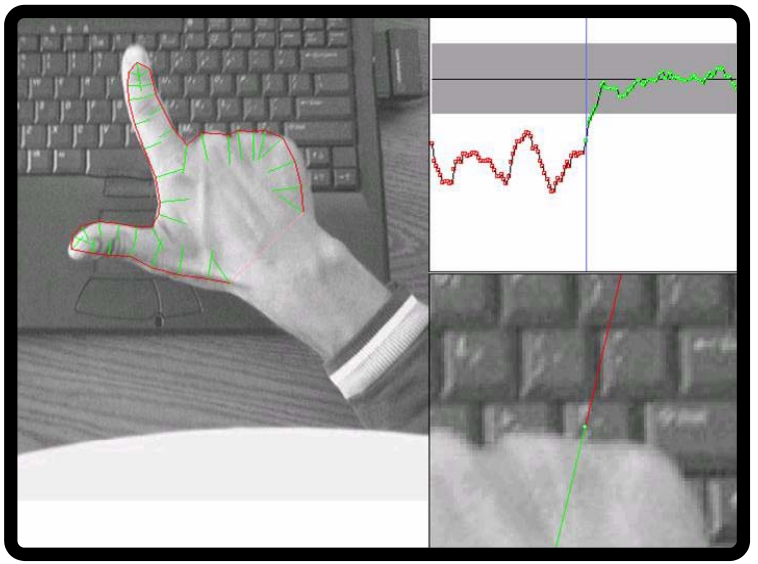


3. Hand detection, tracking, and recognition

Robust single-view detection







View-dependent posture recognition

Hand tracking demo

4. Recognizing body gestures and activity

- Current: Real-time tracking for
  - Interactive digital art applications
  - Autonomous aircraft on carrier flight deck

Restricted EM algorithm for skin classification
Head and hand/arm tracking