Atelier: Acquisition du geste humain par vision artificielle et applications
mardi 27 janvier 2004

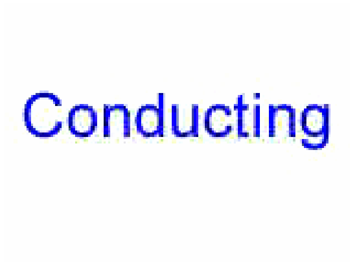
Why Gesture?

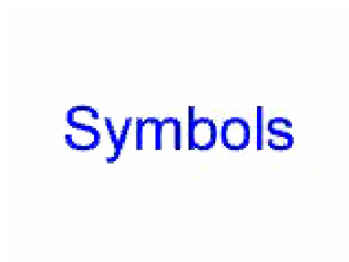
Matthew Turk
Computer Science Department and
Media Arts and Technology Program
University of California, Santa Barbara

http://www.cs.ucsb.edu/~mturk

UCSB

Simple gesture examples





“Cymbals”

Minority Report

US Navy unmanned aircraft scenario

Outline

- ✓ **Why gesture? The Big Picture**
- ✓ **Gesture in the interface**
- ✓ **Related research at UCSB**

Why Gesture? The Big Picture

- Goal: To challenge the gesture research community to think clearly and deeply about
  - What we are doing
  - Why we are doing it
  - How we should do it
  - What we expect to accomplish
- We need go get beyond the "first paragraph syndrome"

Is gesture recognition important?

- Clearly, there are useful applications of gesture recognition
  - Remote control of devices
  - Automatic sign language interpretation
  - Movement characterization for sports, dance, theatre, etc.
  - Human-computer (and human-robot) communication
  - Accessibility, games… and many more
- Computer vision isn't always required
  - Why not use a device (carried or worn)?
  - Is vision a necessity or a preference?
  - Is computer vision just our "hammer"?

Killer app?

- Is there a "killer app" for vision-based gesture recognition?
  - An application that will financially drive and justify extensive research and development in automatic gesture analysis
  - Fills a critical void or creates a need for a new technology

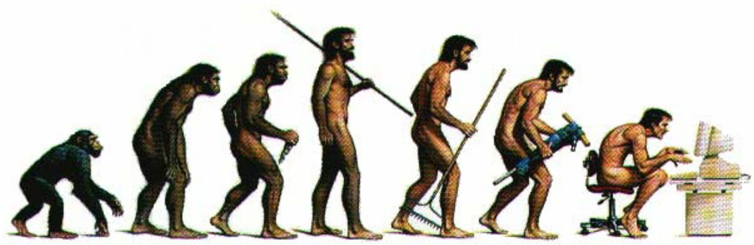
?

No killer app!

- My view: There is no killer app for vision-based gesture recognition
- There are, however, many practical uses of gesture recognition
  - Some vision only
  - Some with other modalities/devices
  - Many that combine modalities
- This is good!!
  - It gives us the opportunity to do the **right thing**
    - **The science of gesture**

Aside: The history of computing

- Computing did not start with the ENIAC, or the Z3, or the Mark I, or the Whirlwind, or the Univac I, or even with Alan Turing
- Computers have a very long history
  - The ideas didn't just appear out of nowhere in the 1940s

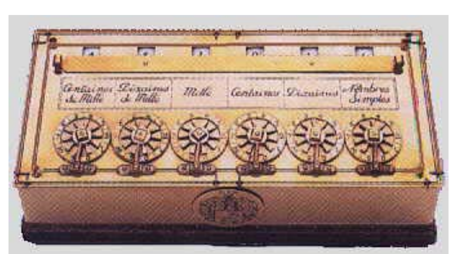


People have always needed computing devices

Computing

- Why did people need to compute?
  - Commerce
  - Astronomy
  - Navigation
  - Warfare
  - Science
  - Taxes
- What is a “computer”?
  - Originally, a job description: ***"a person who computes"***
  - The earliest known reference to “computers”: in 1398 from a writer called Trevisa, who wrote about people who occupied themselves with calculations of time:
  - Boring, repetitive, error-prone calculation of tables!

Blaise Pascal (1623-1662)



1643 – Mechanical adding machine (the "Pascaline")

Gottfried Wilhelm Leibniz (1646-1716)

- Built a better calculating machine
- Dreamed of a universal mathematical language to encode knowledge, and rules to embody logical reasoning
- And of building machines capable of carrying out calculations, freeing the mind for creative thought
- In his vision, "men of good will" would sit around a table to solve a critical problem

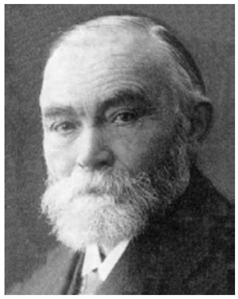
  – Then... ***"Let us calculate!"***
  – The men would calculate and reach a solution, whose correctness would necessarily be accepted by all

George Boole (1815-1864)

- More than 100 years after Leibniz, he didn't know about Leibniz, but proceeded to bring to life part of Leibniz' dream
- Boole's insight: Logical relationships are expressible as a kind of algebra
  – Letters represent classes (rather than numbers)
  – So logic can be viewed as a form of mathematics
- Published ***The Laws of Thought***

- Boole extended Aristotle's simple 3-term syllogisms to a broader range of reasoning
  – Boole: Propositional logic

Gottlob Frege (1848-1925)



- Frege provided the first fully developed system of logic that encompassed all of the deductive reasoning in ordinary mathematics.
- He intended for logic to be the ***foundation*** of mathematics – all of mathematics could be based on, and derived from, logic
- In 1879 he published ***Begriffsschrift***, subtitled "*A formula language, modeled upon that of arithmetic, for pure thought*"

- This was the first example of a formal, artificial language constructed with a precise syntax (rules of grammar)
- As such, the ***Begriffsschrift*** can be considered the ancestor/mother of all current computer programming languages

Georg Cantor (1845-1918)

- The problem of the nature of the infinite had perplexed mathematicians, philosophers, and theologians since Aristotle
  - Is "infinity" only a matter of speaking ("potential infinity") or an actual quantity ("completed infinity")?
- Cantor – against the conventional wisdom of the day, and against significant opposition – accepted the challenge to create a coherent mathematical theory of the actual infinite

- The ensuing debate and disputes over this would eventually lead to key insights into the development of all-purpose digital computers

David Hilbert (1862-1943)

- A brilliant mathematician who was profoundly interested in the foundations of mathematics
- Hilbert cleaned up and filled holes in Euclid's classic treatment of geometry
- In 1900, he presented mathematicians with a grand challenge for the new century: 23 fundamental unsolved problems in mathematics – difficult and important – that set the agenda for generations of mathematicians
  - #1 Cantor's Continuum Hypothesis (Cohen)
  - #2 the problem of the consistency of arithmetic and logic (Gödel)
  - #10 universal solution of Diophantine equations (Matiyasevitch)
- More than a collection of problems – Hilbert's philosophy of mathematics

Kurt Gödel (1906-1978)

- Gödel as a young man was part of the Vienna Circle – a group of philosophers, mathematicians and scientists in the 1920s who founded logical positivism
  - An important goal of philosophy is to develop and study symbolic systems of logic, encompassing mathematics and empirical science
- B. Russell showed that all of mathematics can be encapsulated in a formal logic system
- Gödel chose Hilbert's question of completeness for his doctoral dissertation, and used Cantor's non-finitary methods to prove completeness of Frege's logic
  - Gödel's Completeness Theorem
- Hilbert had previously shown that geometry was ***consistent*** if the arithmetic of real numbers was consistent. Gödel set out to prove this.

Alan Turing (1912-1954)

- Gödel showed that Frege's first-order logic is complete
  - All true FOL statements can be generated
- Turing began to think about Hilbert's "FOL decidability problem"
  - Given some premises and a proposed conclusion, determine whether or not the conclusion can be derived by Frege's rules
  - If this algorithm exists, then all human deductive reasoning can be reduced to "brute calculation"!
    - Fulfilling Leibniz' dream
- Turing thought it might be possible to prove that no such algorithm exists

- Turing proved this... And as a byproduct, he found a mathematical model of an ***all-purpose computing machine.***

Summary: history of computing

- The history of computers is, in large part, a history of ideas
  - "Thinkers" – Visionaries in logic and mathematics
  - "Builders" – Created real machines implementing the ideas

- Turing's vision of human-like intelligence in machines goes back to Leibniz' clear vision of mechanizing human reasoning
  - ***"Let us calculate!"***

- All the great thinkers were motivated not just by current needs and problems, but by vision – ***what could be***

Back to gesture recognition…

- So are you a “thinker” or a “builder”? Or both?

- Thinkers contribute to the theory, the understanding
- Builders leverage the theory, the understanding, to create really useful devices/methods/algorithms

- Of course, thinkers can be builders too, but their building is primarily a tool to aid their thinking!
- And, of course, builders can be thinkers too (one hopes!), but they need to understand the theory behind their building

Another way to view this

- David Marr’s levels of understanding an information processing task:
  - **Theory:** What is being computing and why?
  - **Representation/Algorithm:** How does the computation effect the result?
  - **Implementation:** How is the computation implemented?
- Marr’s warning: Don’t confuse the levels!

Vision Based Gesture Recognition

- Theory level
  - What is a gesture? What is a gesture event?
  - What is the desired output?
  - What is the context of the gesture?
- Representation/Algorithm level
  - What features are important/relevant?
  - How to compute the features?
  - How to represent temporal aspects of gesture?
  - What specifically does recognition entail?
- Implementation
  - Are the sensors adequate?
  - Which classification technique is best? How much training needed?
  - How to initialize? Is it fast enough?

Gesture recognition: State of the art

- My claim is that we have not done a good job in clarifying the "thinking vs. building" distinction, nor in keeping Marr's levels separate
  - "A real-time HMM-based gesture recognition system for robot control"
- As a result, there is no underlying theory of gesture recognition. We're just applying computer vision and pattern recognition to various, loosely-related tasks.
  - We have a "hammer" so everything looks like "nail"!
- Of course, for practical and limited problems, this is all fine.

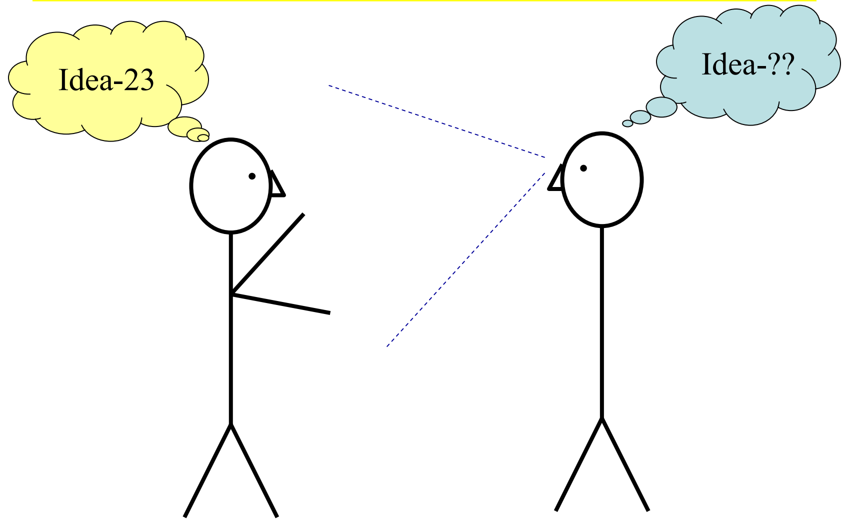
My unsolicited advice

- For the builders:
  - Develop, prototype, and build practical gesture recognition systems for specific applications.
  - Be very concerned about:
    - Scalability
    - Robustness (The Inigo Montoya problem: “You keep using that word. I do not think it means what you think it means.”)
    - Speed
    - Don’t be married to the “vision only” approach
- For the thinkers:
  - Do good science that will lay the groundwork for future generations of researchers

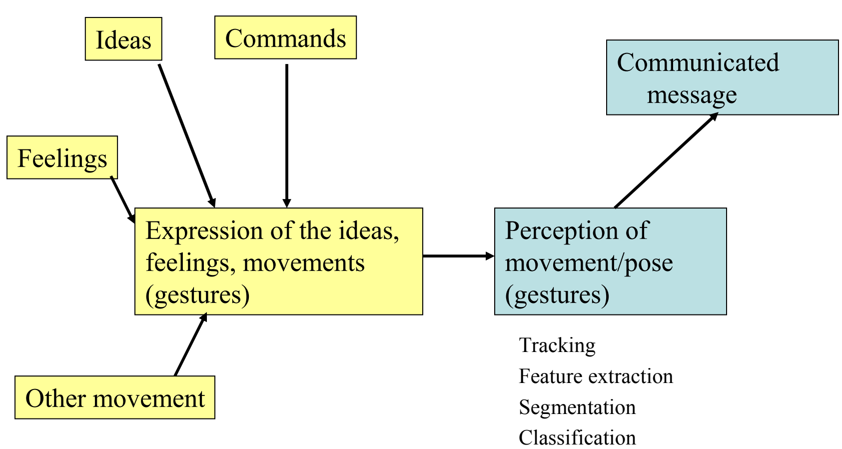
Theory

- What does it mean to do gesture recognition?
  - Just classification? (“Gesture #32 just occurred”)
  - Semantic interpretation? (“He is waving goodbye”)
- What is a gesture?
  - Blinking? Scratching your chin? Jumping up and down? Smiling? Skipping?
- What is the purpose of gesture?
  - Communication? Getting rid of an itch? Expressing feelings?
- What is the context of gesture?
  - A conversation? Signaling? General feedback? Control?
  - How does context affect the recognition process?

Gesture recognition



Gesture recognition



What is gesture?

- What is gesture?
  - "body language"
  - "silent language"
  - "visual shortcuts"
- Classes of gesture
  - Instinctive (e.g., baby smile/laugh)
  - Coded (e.g., sign language)
  - Acquired (day-to-day, social)
- Gestures can be warm, menacing, instructive, sensuous, unintentional…

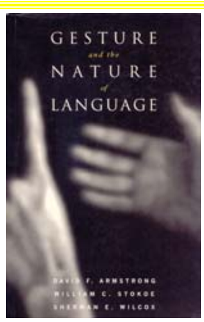
Human gesture

- Humans can produce up to 700,000 different physical signs (M. Pei)
- 250,000 facial expressions (Birdwhistell)
- 5000 hand gestures (Krout)
- People gesture when talking on the phone(!)
- Blind people may gesture when talking to each other
- Gestures often have cultural specific meanings
  - Don't hitchhike with your thumb out in Nigeria
  - Be careful how you cross you legs in some places
  - Don't flash the "victory" sign everywhere
  - Etc….

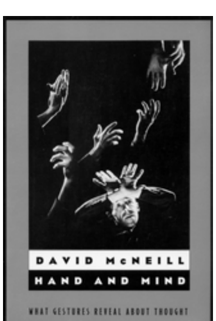
Who studies gesture?

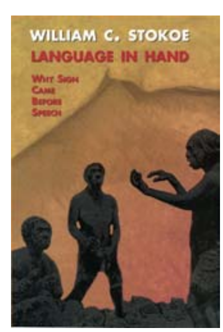
- Linguists, anthropologists, sociologists, psychologists, cognitive scientists, kinesicists, etc. – all study gesture (as it relates to thought, language, and speech)
  - And all have different viewpoints and different theories, and much debate
  - (Kinesicists do kinesics – the study of body motion)
- Charles Darwin had a lot to say about human and animal gestures (including facial expressions)
- Desmond Morris (British social anthropologist)
  - Very influential, studied human and animal behavior
  - Manwatching (1977), Gestures (1979)

Recent influential research



- Adam Kendon
  - Univ. of Pennsylvania
- David McNeill
  - Univ. of Chicago
- …and several others

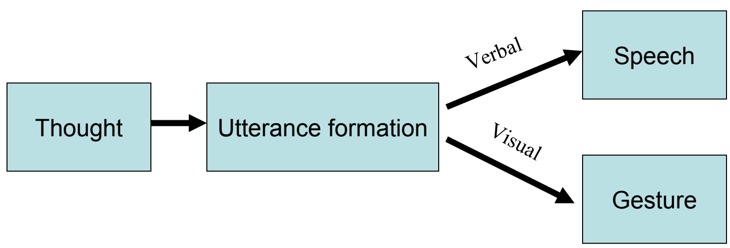




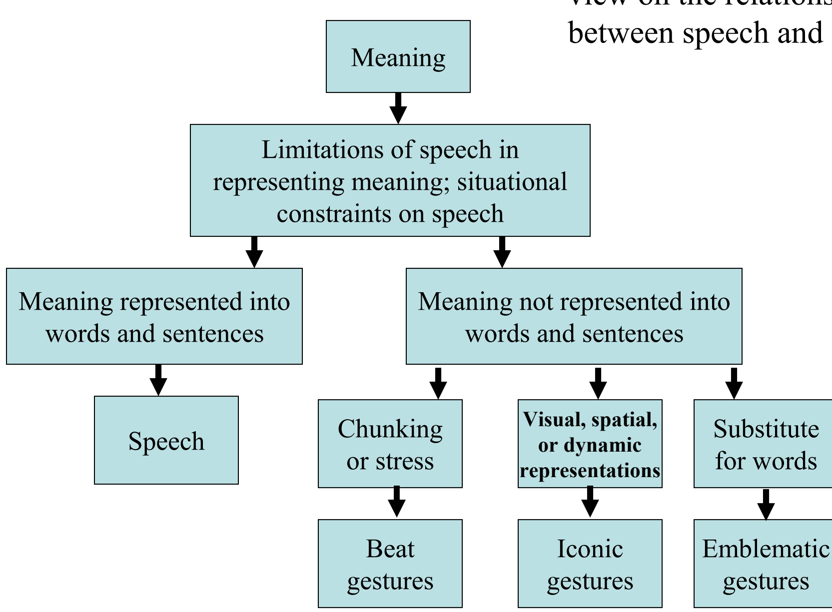
Gesture and thought

- Clearly, gesture is related to thought – but how?
  - Is gesture “rendered thought”?
  - Is gesture its own language?

- How is gesture related to verbal speech?
  - One view:



A summary of Kendon’s view on the relationship between speech and gesture



Gesture and thought

- Gesture is intertwined with thought and language – all part of one system
  - Gesture is not just "visible speech"
- Gesture and thought impact one another
  - Gesture is not just rendered thought
  - Neither is language

McNeill's gesture taxonomy

- Iconic
  - Pictoral gestures corresponding to a concrete object or event
  - The gesture resembles (presents a picture of) the coincident part of the utterance. Both speech and gesture parts are important.
- Metaphoric
  - Pictoral gestures corresponding to an abstract idea
- Beat
  - Nonpictoral; underscores the rhythmic pulsation of speech
- Diectic
  - Pointing gestures

McNeill's gesture taxonomy

- Iconic
  - Pictoral gestures corre
  - The gesture resembles
    part of the utterance. E
    important.
- Metaphoric
  - Pictoral gestures corresp
- Beat
  - Nonpictoral; undersco
- Diectic
  - Pointing gestures

Typically have three stages:
- Preparation
- Stroke
- Retraction

Two stages: in/out, up/down, right/left…

McNeill on (spontaneous) gesture and language

Evidence that speech and gesture are two sides of a single process of constructing and presenting meanings:

- Gestures occur only during speech
  - Actually, listeners do gesture…
- Gestures and speech are semantically and pragmatically co-expressive
  - They present the same or closely related meanings and functions
- Gestures and speech are synchronous
- Gestures and speech develop together in children
- Gestures and speech break down together in aphasia

Which came first?

- Many believe that language originated in gesture

- Some argue that speech ***is*** gesture
  - Speech: vocal-articulatory gestures
  - Gesture: visible gestures

Context

- *Context* underlies the relationship between gesture and meaning
- Except in limited special cases, we can't understand gesture (derive meaning) apart from its context
- We need to understand both gesture **production** and gesture **recognition** together (not individually)
- That is, "gesture recognition" research by itself is, in the long run, a dead end

So…

- Gesture recognition is not just a technical problem in Computer Science
- A multidisciplinary approach is vital to truly “solve” gesture recognition – to understand it deeply
  - “Thinkers” and “builders” need to work together
- Don’t believe the theories!
  - At least, don’t take them too seriously…

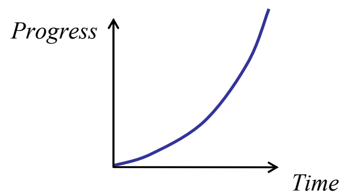
Outline

- ✓ Why gesture? The Big Picture
- ✓ **Gesture in the interface**
- ✓ Related research at UCSB

Observation

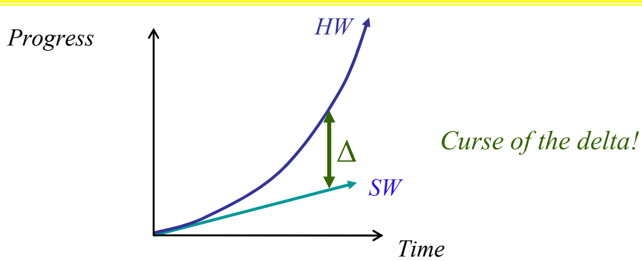
- **Moore's Law** has driven computer technology for decades

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

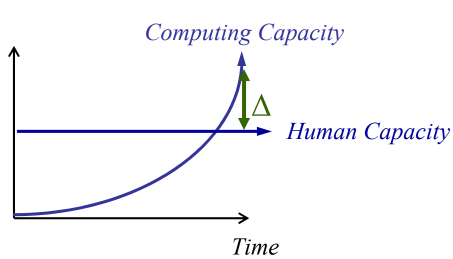


- But… there has been no Moore's Law for user interfaces!
  - The result?

The result



Another view:
There's no Moore's Law for people!



Curse of the delta

What to do?

- Maybe we need to rethink the way we interact with computers
- Question: What's the ultimate user interface?
  a) A well-designed machine/instrument
  b) An assistant or butler
  c) None! UIs are a necessary evil
  d) All of the above

- UI Goals:
  – Transparency
  – Minimal cognitive load
  – Task-oriented, not technology-oriented
  – Ease of learning, ease of use (adaptive)

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

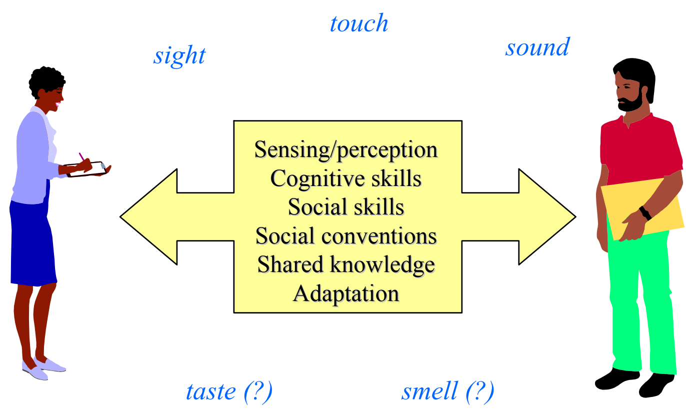
Highly interactive, multimodal interfaces modeled after natural human-to-human interaction

- 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

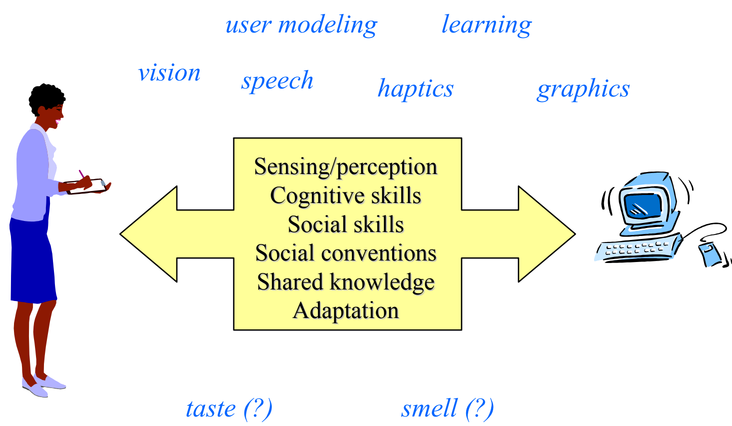
Not just passive

Multiple modalities, not just mouse, keyboard, monitor

Natural human interaction



Perceptual Interface



Early example

"Put That There" (Bolt 1980)…

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 – multiple users, multiple tasks
  - Transfer of natural, social skills – easy to learn
  - People already anthropomorphize technology
    - E.g., Reeves & Nass studies
  - 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

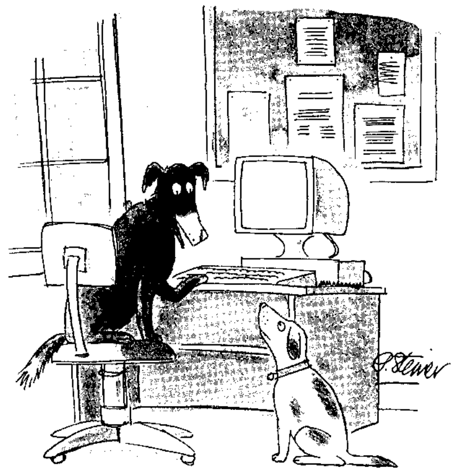
Haptic I/O
Affective computing
Tangible interfaces
Sound recognition
Sound generation
User modeling
Conversational interfaces

(input and output)

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 – yipes!
- 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

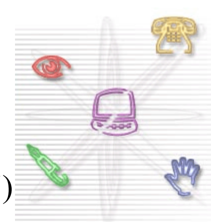
- Technology (the easy one)
  – Lots of researchers worldwide
  – Increasing interest
  – Consistent progress

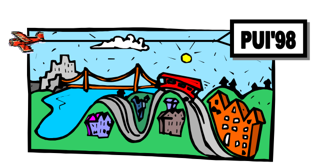
- The Marketplace (the hard one)
  – But there's growing convergence: hw/sw advances, commercial interest in biometrics, accessibility, recognition technologies, virtual reality, entertainment….

PUI, ICMI

PUI Workshop (1997, 1998, 2001)
http://www.cs.ucsb.edu/PUI

ICMI (1996, 1999, 2000, 2002, 2003, 2004, 2005)
http://icmi.cs.ucsb.edu





Outline

- ✓ Why gesture? The Big Picture
- ✓ Gesture in the interface
- ✓ **Related research at UCSB**

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
  - ….

UCSB "Four Eyes" Lab

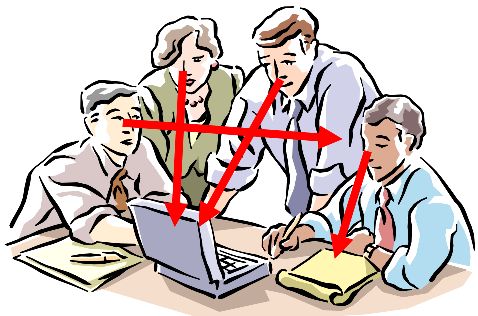
- People:
  - CS Faculty: Matthew Turk (2000) and Tobias Hollerer (2003)
  - 1 faculty visitor, 8 PhD students, 2 MS students, 2-4 BS students
- Current funding:
  - NSF ITR, NSF IGERT, U.S. Navy, Lawrence Livermore Laboratory, UCSB Research Across Disciplines, …
- Collaborations at UCSB:
  - Psychology, Geography, Media Arts and Technology, Electrical and Computer Engineering, Cognitive Science, Center for Information Technology and Society, ….

UCSB "Four Eyes" Lab

- Current projects (Turk)
  - Face tracking
  - Facial expression analysis
  - Hand detection/tracking/postures
  - Body gestures, activity
  - Continuous multimodal biometrics
  - Immersive environments (VR)

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)

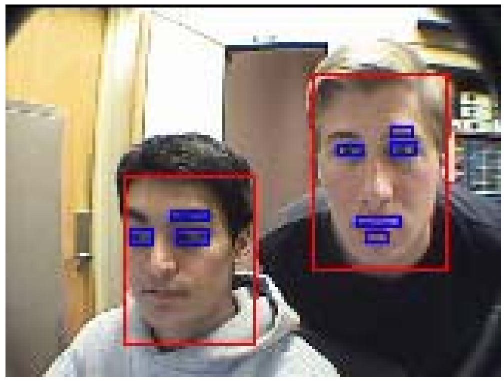


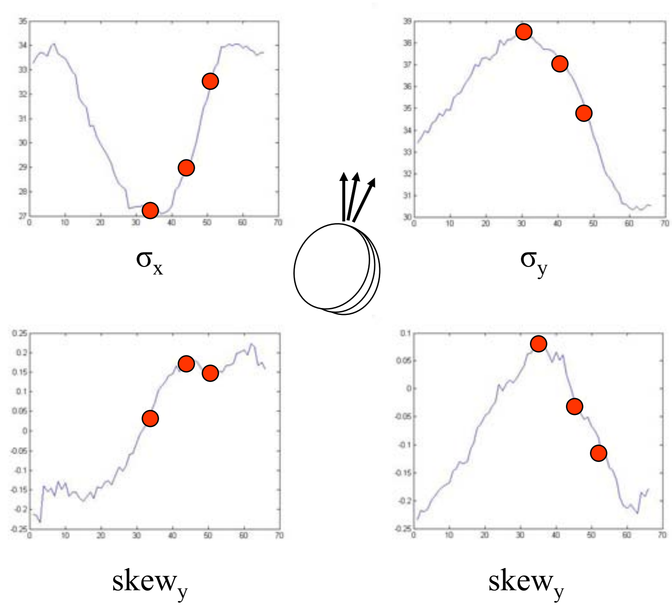
Action understanding
Meeting annotation
Audience feedback
Videoconferencing
Etc.

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)
  - $f\,(\textit{statistical measures}) = \textit{direction}$

Example results

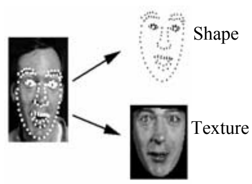




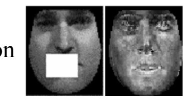
2. Active Wavelet Networks for Face Alignment

- Facial feature detection and tracking
- Goal: Accurate face pose tracking to support VBI applications

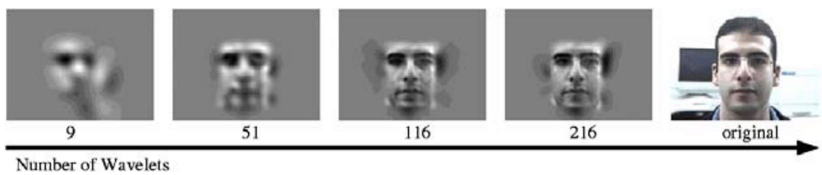
AAM



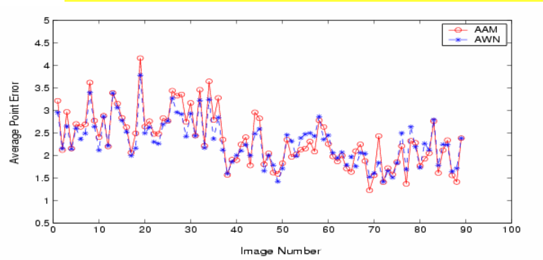
Problem: Partial Occlusion



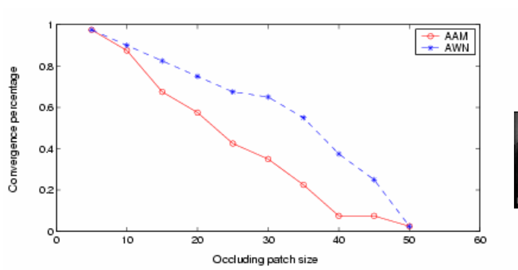
Main idea: Replace AAM *texture* model by a Gabor wavelet network



Active Wavelet Networks (AWN)

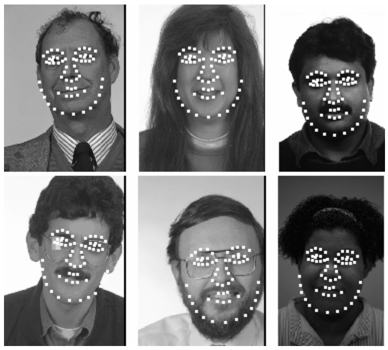


Similar performance to AAM in images under normal conditions.



More robust against partial occlusions.

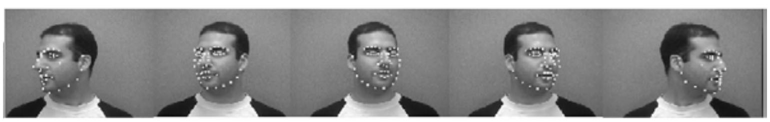
Face Alignment with AWNs



Using 9 wavelets, the system requires only 3 ms per iteration.

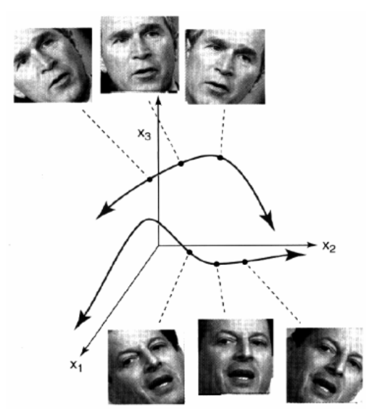
In general, at most 10 iterations are sufficiently for good convergence (PIV 1.6Ghz).

We have extended AWN to **multi-view AWN** for real time face alignment under large pose variation



3. Facial expression analysis

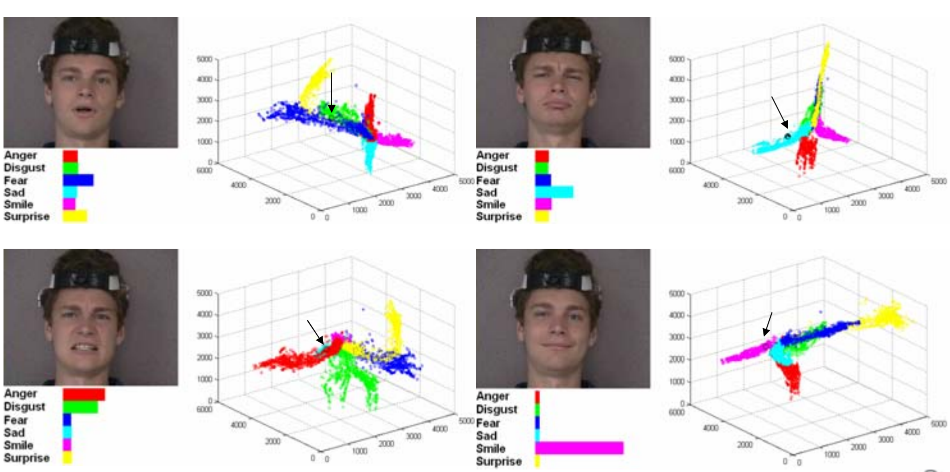
- Facial expression representation and visualization
- Use non-linear manifolds to represent dynamic facial expressions
- Intuition:
  - The images of all facial expressions by a person makes a smooth manifold in (high-dimensional) image space, with the "neutral" face as the central reference point



Probabilistic expression analysis on manifolds

- In the embedded space, basic emotional expressions become paths on the manifold, emanating from the center (neutral expression). Blends of expressions will lie between those curves.
- Each path consists of several clusters. A probabilistic model of transition between the clusters and paths is learned through training videos.
- The transition between different expressions is represented as the evolution of the posterior probability of the six basic paths

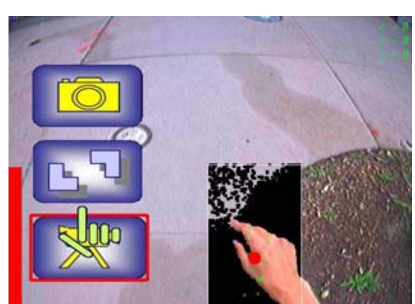
Manifold visualization of expression



Example

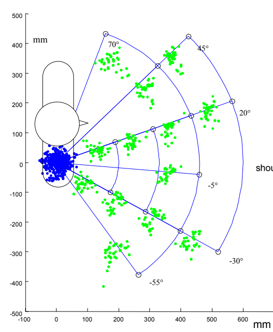
4. Hand Gesture Interfaces for AR

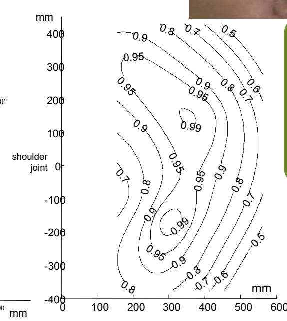
- Goal: To build highly robust CV methods that allow out-of-the-box use of hand gestures as an interface modality for VR and AR

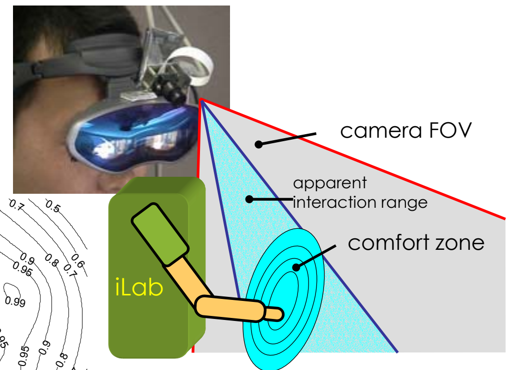


How to design gestural interfaces?

Comfort Zone: “Reaching postures assumed voluntarily, despite the availability of compensatory postures”

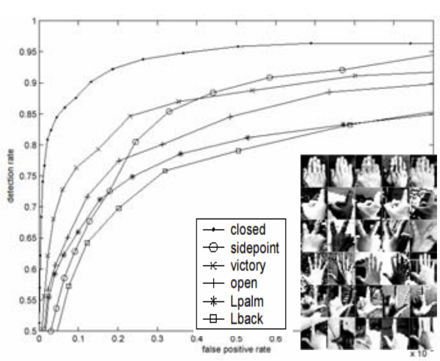


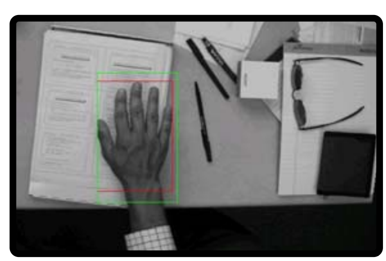


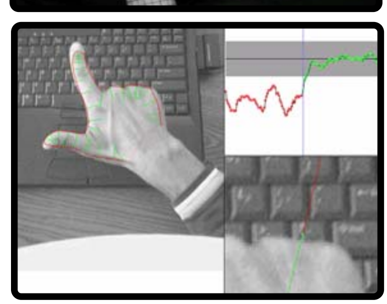


Hand detection, tracking, and recognition

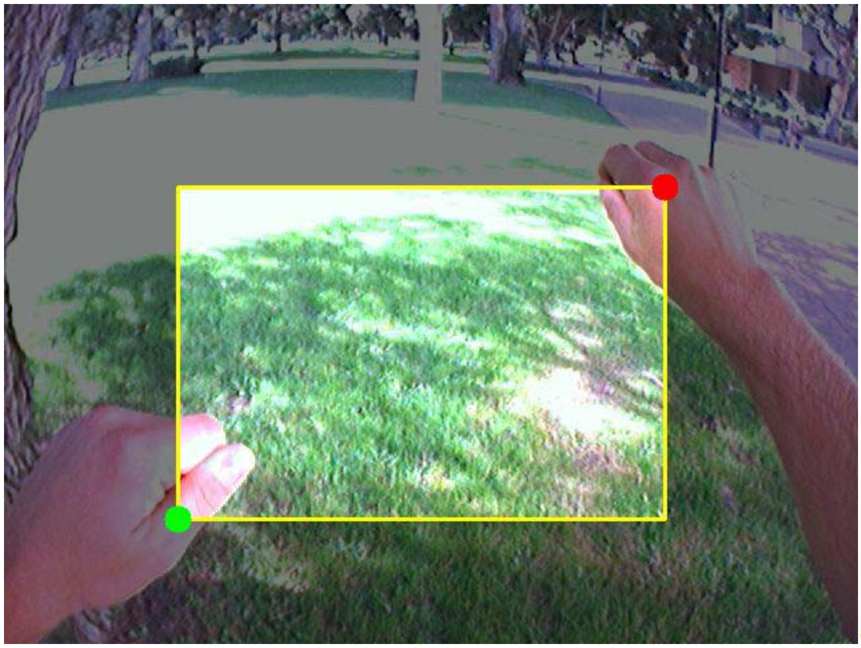
Robust single-view detection





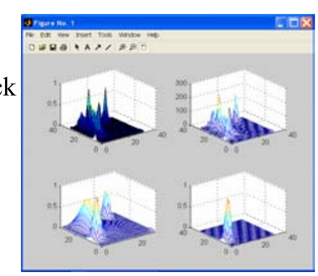


View-dependent posture recognition

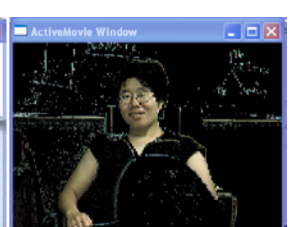


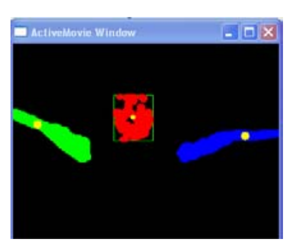
5. Recognizing body gestures and activity

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



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

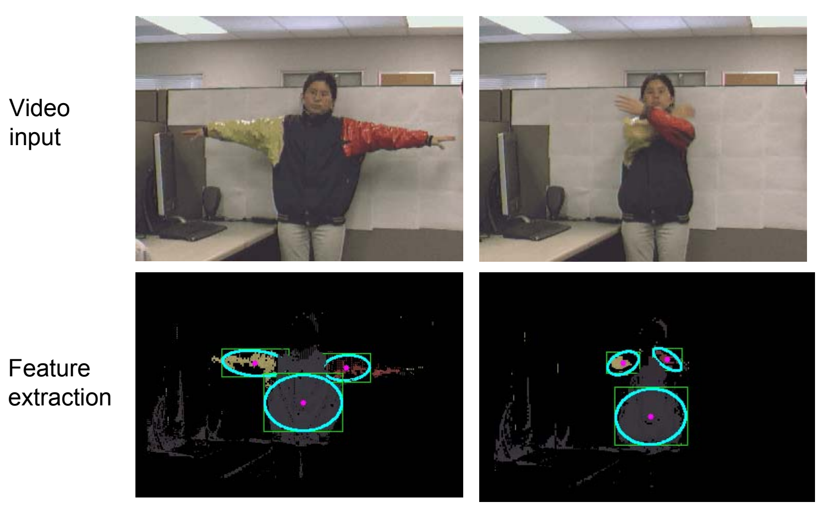




Application: Autonomous unmanned aircraft

- Deck handling – automatic recognition of gestures while taxiing on the deck of an aircraft carrier

Feature Extraction

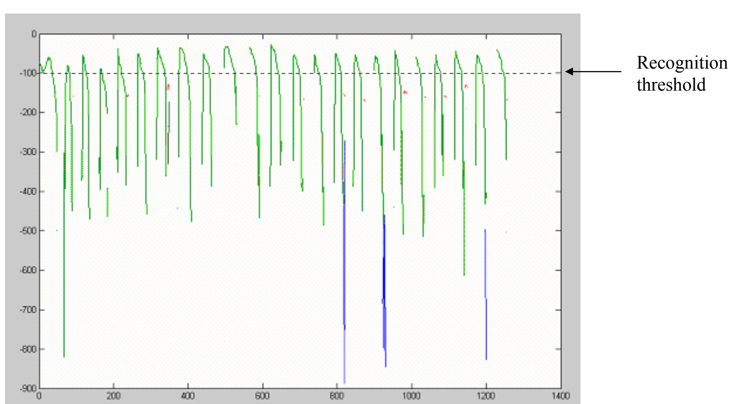


Early recognition results

Recognition results (%)		Output			
		Fold wing	Spread wing	Move back	Turn right
Input	Fold wing	100	0	0	0
	Spread wing	0	100	0	0
	Move back	0	0	100	0
	Turn right	22	0	0	78

Confusion matrix from experiment with four gestures

Early recognition results



Input: Video sequence of “move back” gesture repeated 23 times

Green is log-likelihood of “move back” – other colors depict other gestures.
The green peaks show successful recognition of “move back” 23 times.

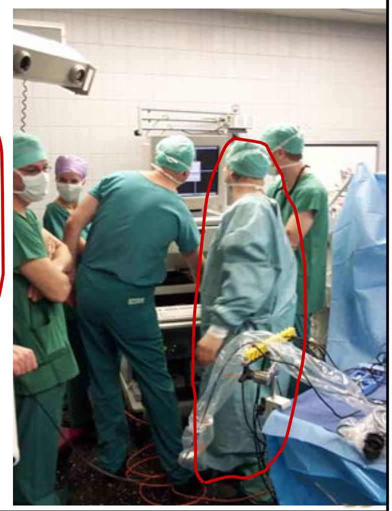
It's very difficult!

- We successfully recognized gestures from continuous video
  – Gestures: *Fold wings, spread wings, move back, turn left, turn right*
- Experiments were in a lab environment
  – Much more benign than the typical UAV deck environment
- Color based low-level features will eventually be replaced with specially enhanced features
  – "Vision-only" approach is not feasible in near-term
- Preliminary results of gesture recognition using HMMs validate the approach and clarify the path to eventual application and deployment

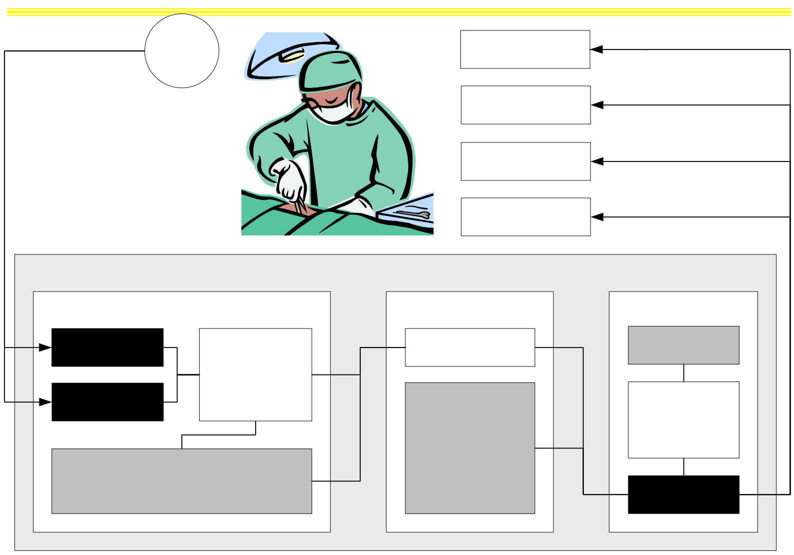
Surgeon-computer interface

- Work by Sebastien Grange, EPFL
- Uses depth data (stereo camera) and video

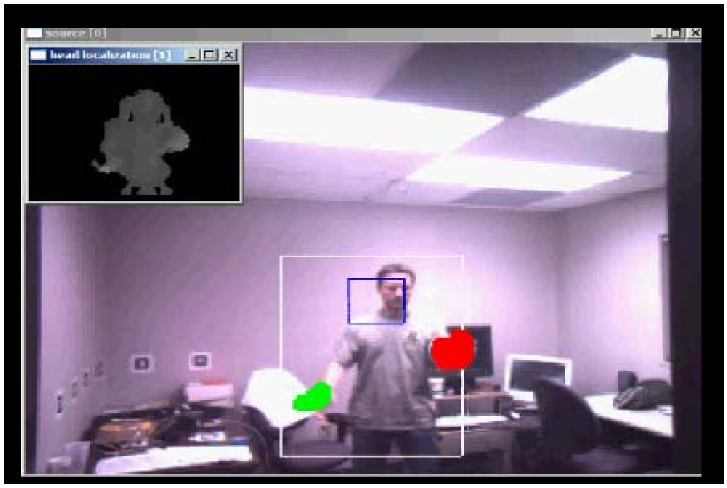
The Operating Room



System diagram



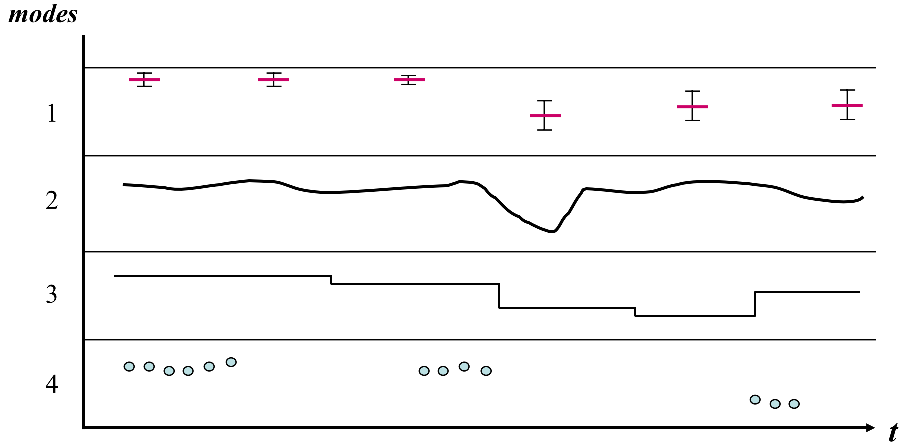
Initial tracking and classification example



6. Continuous multimodal biometrics

- Multimodal – Using multiple modalities (sensor channels) and integrating their output
  - Higher accuracy, improved usability, applicable to a wider variety of people and environments…
- Continuous – Ongoing evaluation rather than one-time, point of access (POA) verification
  - Continue to sense and evaluate – integrate over time as well as across modalities
  - Different modes have different temporal characteristics, reliabilities, and "sweet spots"
  - Gesture (gait, etc.) can be used as a biometric!

Horizontal and vertical integration



Integrate across modes/channels; integrate across time

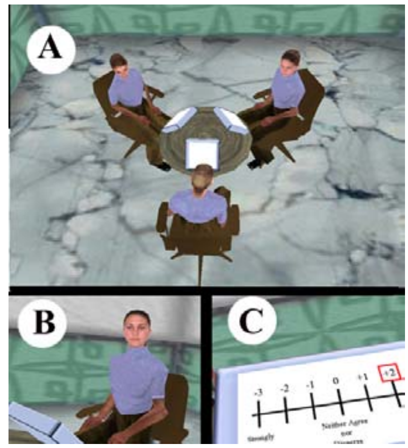
What's possible with this?

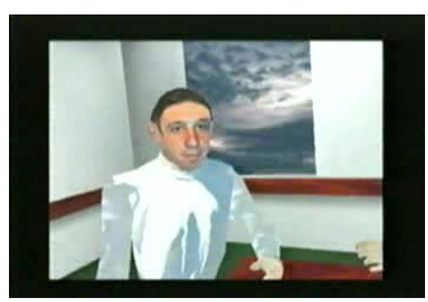
- Continuous authentication instead of once-and-for-all access
- Multiple levels of authentication, e.g.:
  - Selectively remove access to data
  - Ask for re-verification
  - Shut down access
  - Alert security personnel
  - Subtly set up a trap

Preliminary results with Bayes classification and temporal propogation of uncertainty

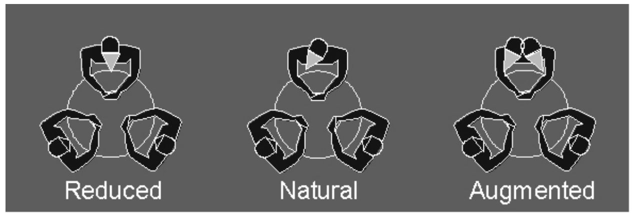
7. Transformed Social Interaction

- What are the effects of manipulating reality in collaborative virtual environments?



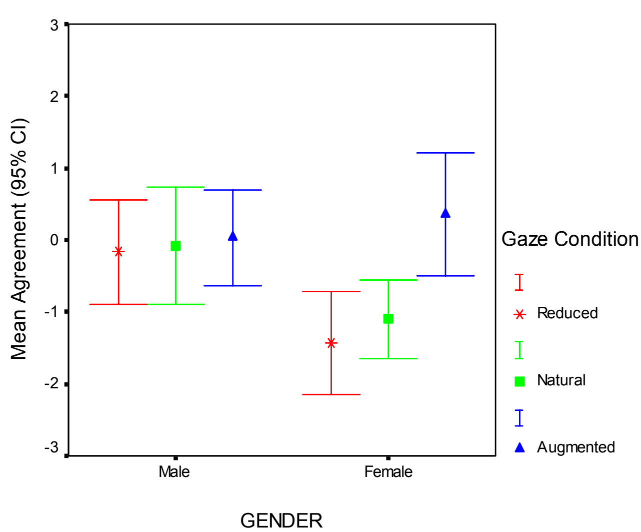


Transformed Social Interaction (cont.)



- Is it possible to increase one's power of <u>persuasion</u> by "augmented non-zero-sum (NZS) gaze"?
  - Presenter gives each participant > 50% of attention

Initial results



Four Eyes Lab – Looking forward

- Technology: More of the same…
  - Fundamental issues in developing robust, real-time, working computer vision technologies
    - Multidisciplinary approach
  - Multimodal integration
    - Speech, sound, language, haptics, user modeling, visualization
  - Main application areas
    - HCI, entertainment, digital art, visualization, …
  - Context, context, context
    - Specific applications
    - General human-human interaction
- Motivation
  - Provide better, more compelling HCI technology in many (all?) environments

Summary

- Why gesture?
  - Good gesture recognition technology will have many uses, and will spur new applications
  - Current efforts are mostly in very limited domains and quite fragile
  - We need a better understanding of gesture (THEORY)
  - Interdisciplinary approach is vital
- No killer app, but broad application in general HCI/PUI
- Come visit us at UCSB and see the demos for yourself!

UCSB Four Eyes Lab

Thanks to: Mathias Kolsch, Changbo Hu, Rogerio Feris, Ya Chang, Haiying Guan, Jeremy Bailenson, Andy Beall, Alphan Altinok