Multimodal User Interfaces 2019

[4] Voice, Gesture, Tangible Interaction

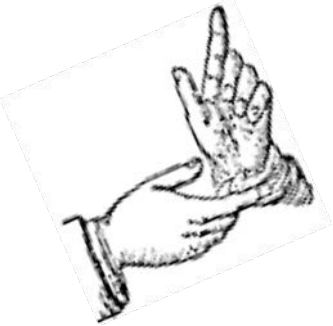
Denis Lalanne

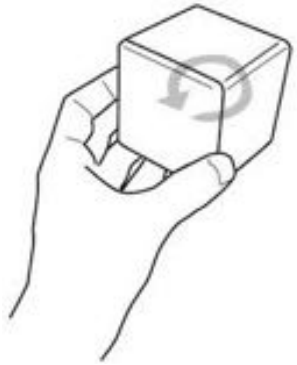
12.03.2019

Thx to Stefano Carrino, Bruno Dumas for their inputs

Agenda

- Voice-based interaction
  - ➢Recognition
    - ✓ Design parameters
  - ➢Synthesis
    - ✓ Design parameters
- Gesture-based interaction
  - ➢Recognition
    - ✓ Design parameters
    - ✓ Gesture recognition approach
- Tangible interaction





[4.1] Voice-based Interaction

Voice-Based Interaction

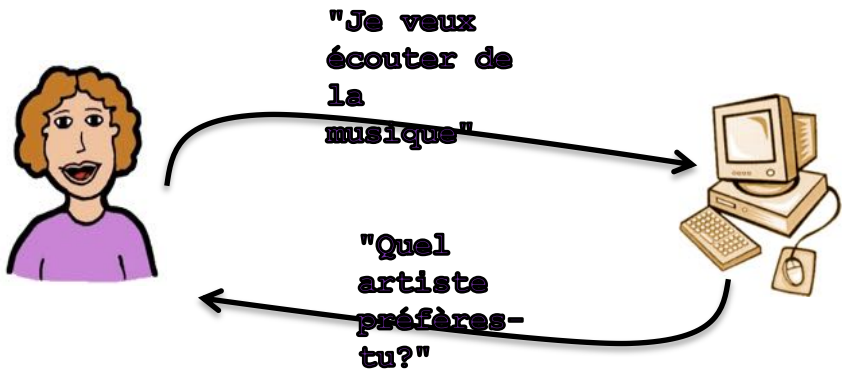
- Why?
  - Speech is a natural way to interact with people
  - Fast
  - If other interactions are not possible
    - ✓ Free-hand interaction

Voice-Based Interaction

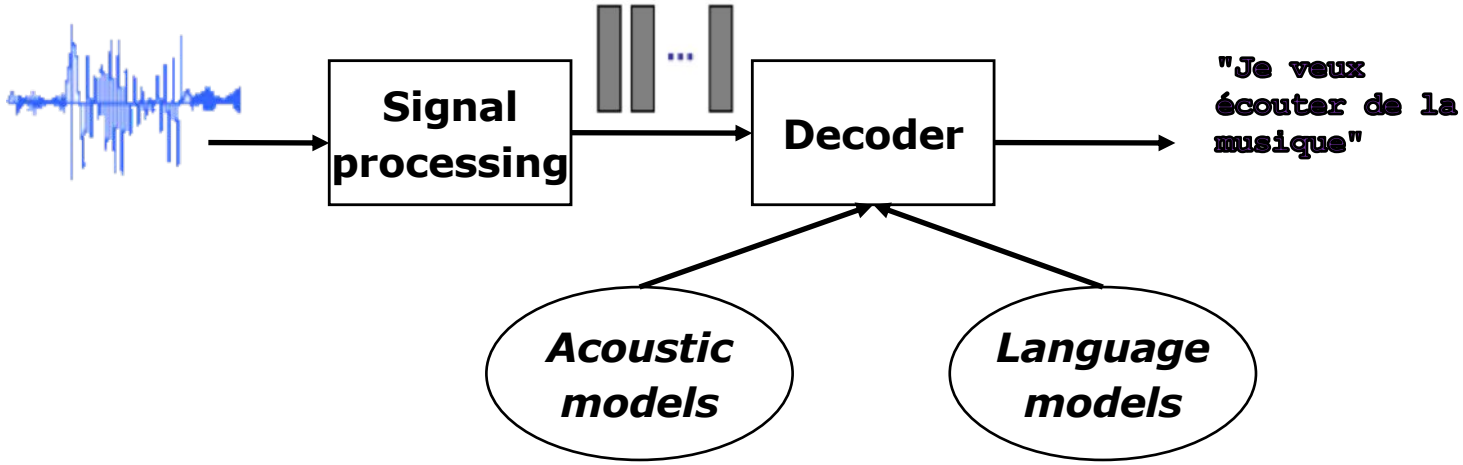
- Voice recognition

- Voice synthesis

- Bi-directional dialog



Voice Recognition



- Signal processing: Reduce dimensionality of signal, noise conditioning
- Decoder: Transcribe speech to words
- Acoustic model: produces phonemes
- Language models: grammar, vocabulary, etc.

Voice recognition – Design parameters

- Speaker-Dependent Vs Speaker-Independent
  - S.D.:
    - ✓ Advantage: better results
    - ✓ Drawback: dedicated to a single user
    - ✓ Must be trained prior to its use
  - S.I.:
    - ✓ Advantage: speaker independent
    - ✓ Drawbacks: increased complexity, lower accuracy, and slower response
  - Speaker Adaptive
- Keyword spotting Vs Continuous speech
  - Isolated-word recognition system have higher performance because of the reduced vocabulary

Voice recognition – Design parameters

- Grammar
  - Used to define which words are valid to the system, and the syntax
  - Set of syntactic and semantic rules, based on task requirements
    - ✓ E.g. Grammar: e.g. voice-operated interface for phone dialing

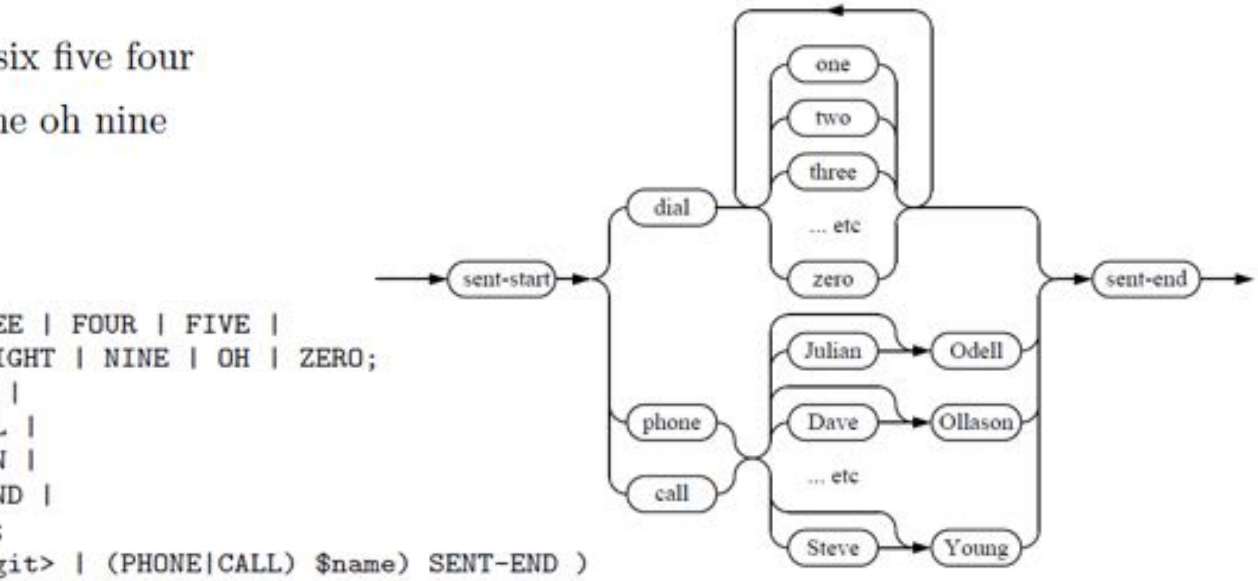
Dial three three two six five four

Dial nine zero four one oh nine

Phone Woodland

Call Steve Young

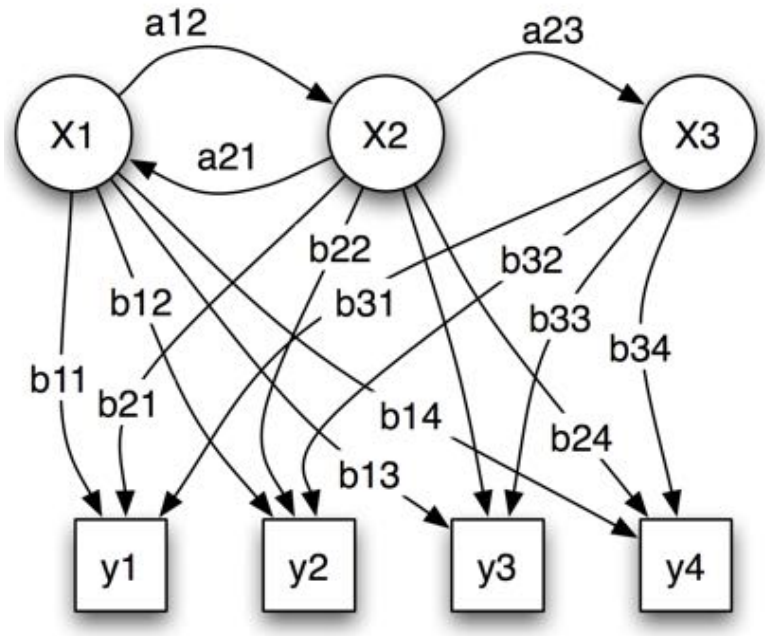
```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name  = [ JOOP ] JANSEN |
         [ JULIAN ] ODELL |
         [ DAVE ] OLLASON |
         [ PHIL ] WOODLAND |
         [ STEVE ] YOUNG;
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )
```



- Vocabulary (small Vs large)
  - Typically task dependent
  - Small vocabularies are easier to recognize

Algorithms – HMM

- Recognition through Hidden Markov Models
  - Statistical, machine learning approach
  - Advantages
    - ✓ User dependent or independent
    - ✓ No training of full system necessary when vocabulary changes
    - ✓ Spotted as well continuous recognition
  - Disadvantages
    - ✓ Generally Slow (training)
    - ✓ The amount of data that is required to train an HMM is very large

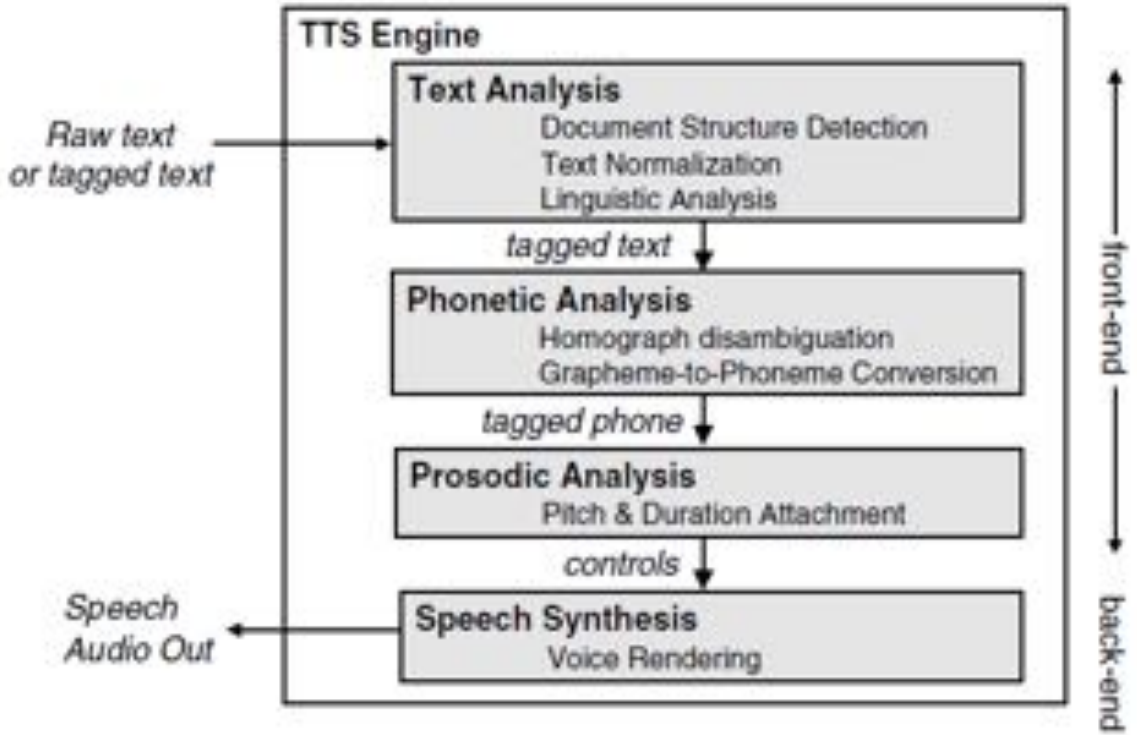


Speech Recognition : Challenges, advantages, drawbacks

- Safety
  - Will an error in recognition have a serious impact on the safety of the final application?
- Three advices for SR based interactions:
  - Train the system in the environment in which it will be used
  - Don't try to use SR for tasks that don't really fit
  - Incorporate error correction mechanisms
- The feedback challenge
  - Hard to provide a continuous feedback
- Noisy environments
  - Bad SNR (Signal to Noise Ratio)
- Privacy issues

Voice Synthesis

- The goal of Text-to-Speech (TTS) synthesis is to convert arbitrary input text to intelligible and natural sounding speech so as to transmit information from a machine to a person.



- Advantage: In cases eyes busy, displays impossible/insufficient, translation, etc.

Voice Synthesis – Challenges

- Front-end
  - Punctuation misinterpretations
  - Abbreviations and acronyms
  - Proper names (test)
- Back-end & system design choices
  - Synthesis by rule
    - ✓ Using experts knowledge
  - Concatenative synthesis
    - ✓ Employ recordings of humans
  - Robotic: simulate vocal cords (video)
- TTS evaluation
  - TTS goal: produce speech from any text as natural and intelligible as human speech
  - An open research topic
  - Accuracy, intelligibility, and naturalness

Tools and libraries

- Tools:
  - Synthesizer:
    - ✓ Reverso: http://www.reverso.net
    - ✓ Google: http://translate.google.com
  - Recognizer:
    - ✓ Youtube captioning
    - ✓ Google voice
    - ✓ Naturally speaking (Nuance)
    - ✓ Siri

- Libraries
  - Speech - input: Sphinx 4 - HTK - W7
  - Speech - input / output: CLSU toolkit - CMU Communicator - Ravenclaw - Nuance - Windows Tools
  - Speech output: FestVox - with 3D face animation : iFace
  - Sound output: Sonification Sandbox
  - Conversational speech system: SpeechActs

[4.2] Gesture-Based Interaction

First Definition of a Gesture

- A motion of the limbs or body made to express or help express thought or to emphasize speech
- The act of moving the limbs or body as an expression of thought or emphasis
- An act or a remark made as a formality or as a sign of intention or attitude
- A succession of postures

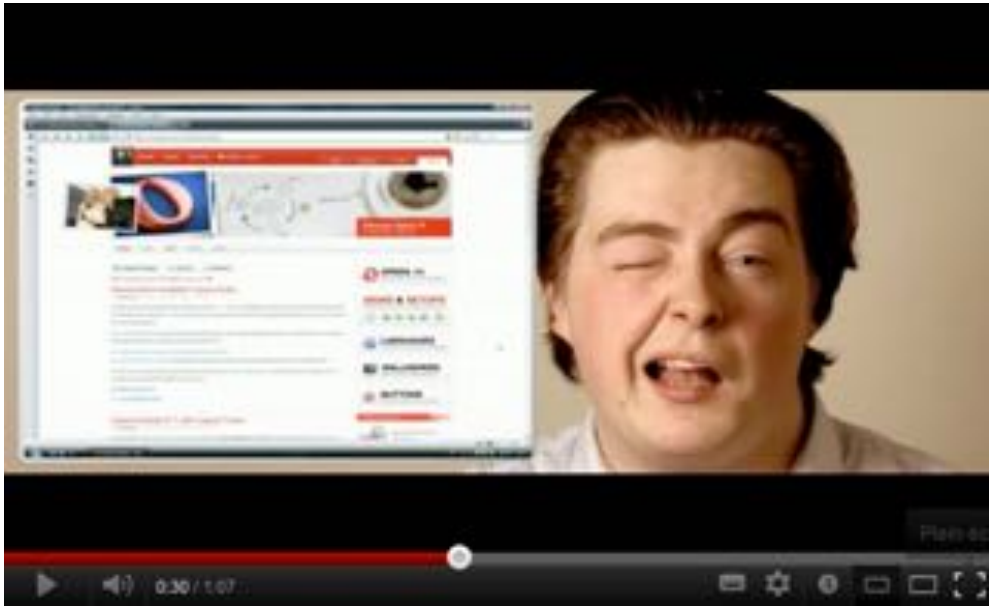
- Our own definition (for this lesson):
  - An **intentional** movement of the body or limbs performed to convey information

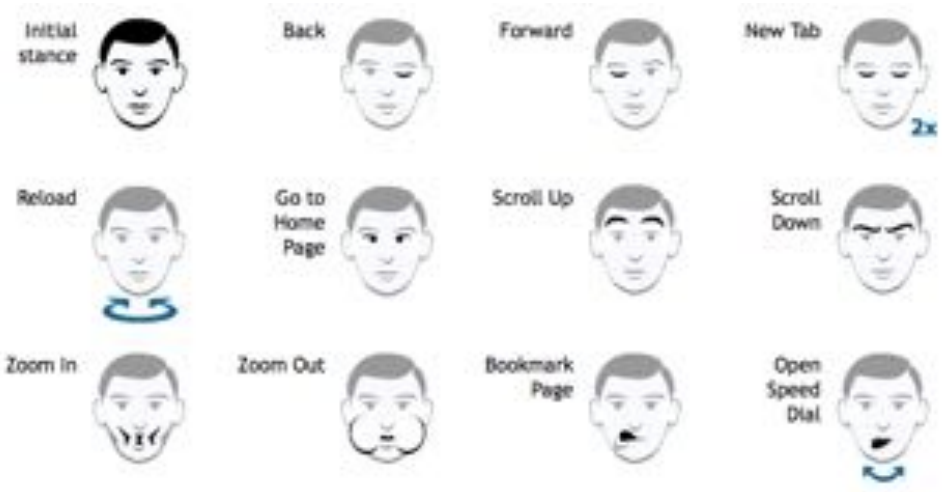
More Formal Definition of a Gesture

A gesture is a form of non-verbal communication in which visible bodily actions communicate particular messages, either in place of speech or together and in parallel with words. Gestures include movement of the hands, face, or other parts of the body.

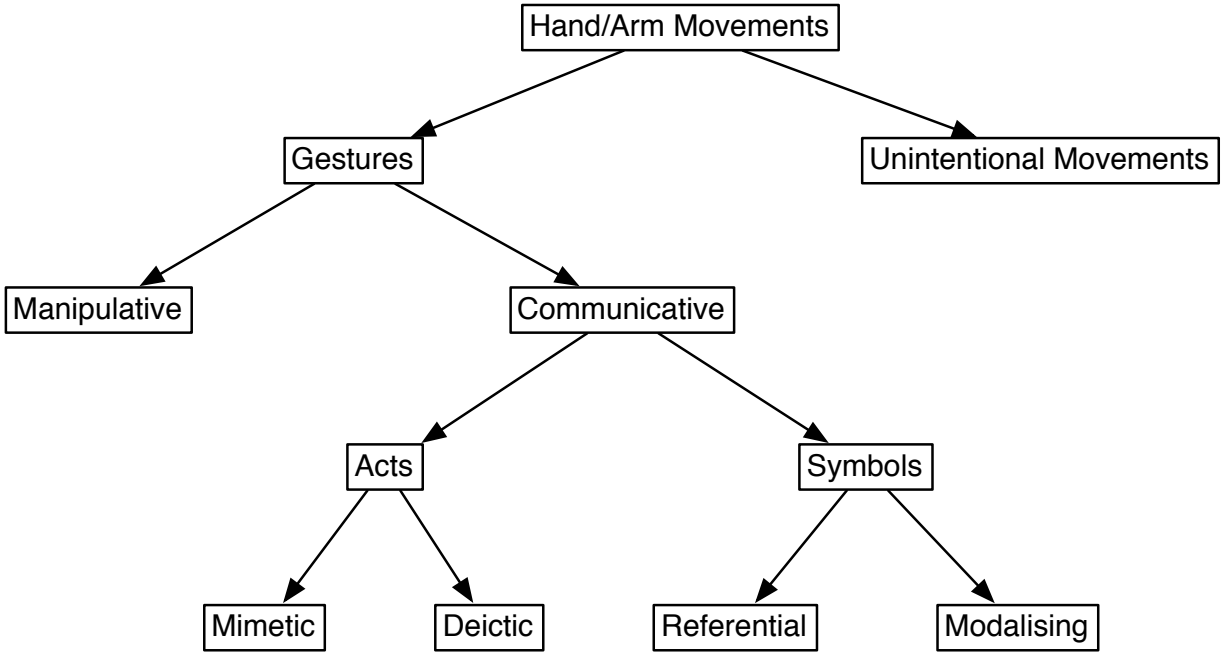
Kendon, Adam. *Gesture: Visible Action as Utterance*. Cambridge University Press (2004).

Are These Gestures?





Taxonomy of Hand/Arm Gestures



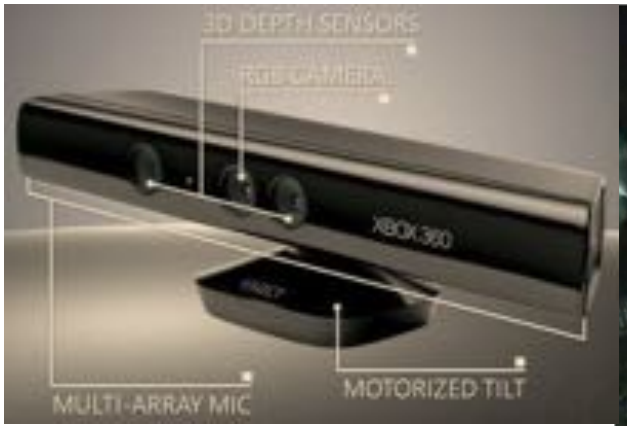
- Gestures vs. unintentional movement (gesticulation)
- Communicative vs. manipulative gestures
  - Manipulative gestures are used to act on objects in an environment (object movement, rotation, etc.)
  - Communicative gestures have an inherent communicational purpose

Gesture Recognition

- Gesticulation vs. Gestures
  - Gesticulation provides also information!
- Postures vs. Dynamic Gestures
  - Postures (aka static gesture, pose)
  - Dynamic gestures: a sequence of postures
- Multi-dimensional gestures
  - 2D gestures
  - 3D gestures

Devices for Gesture acquisition

- Camcorders and webcams
- Skeleton tracking – range imaging – time-of-flight
- Gloves
- Accelerometers
- Multi-touch interfaces





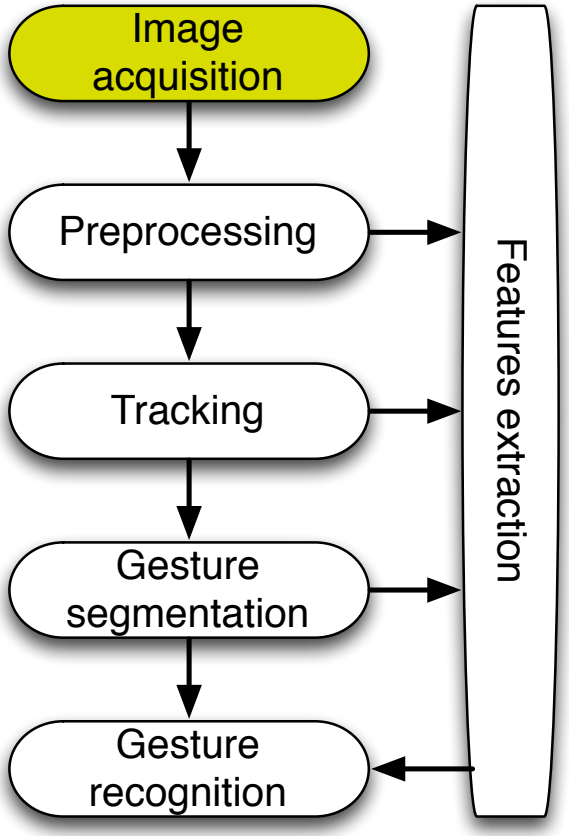
Dynamic Gesture Recognition Processing

- Challenge: get an image captured by a webcam:

- … And track the hand, to recognise a gesture (dynamic)

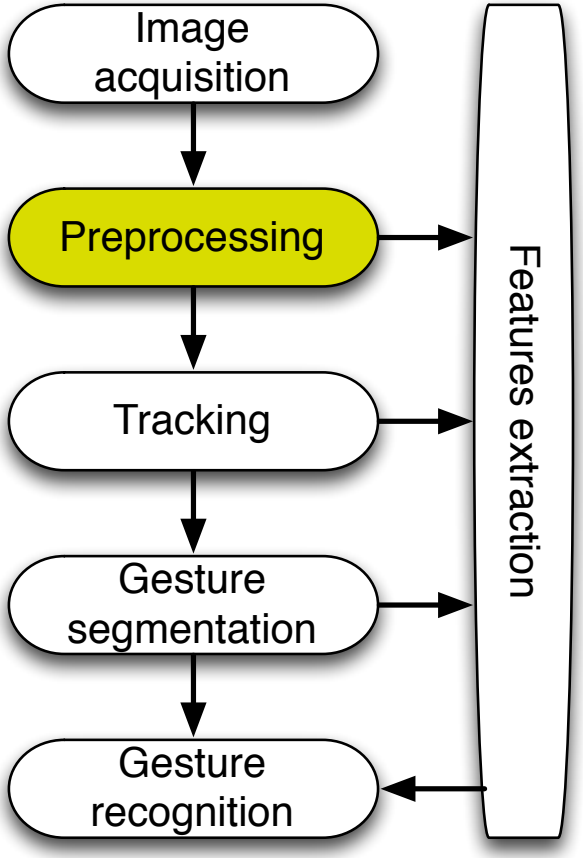
Dynamic Gesture Recognition Processing

- A generic processing pipeline for dynamic gesture recognition
- Especially applicable to computer vision based gesture recognition
  - Applicable to most algorithms mentioned in this lesson, with modifications
  - E.g.: “Image acquisition” would be swapped with “Vectorised data acquisition” in the case of accelerometer based gesture recognition



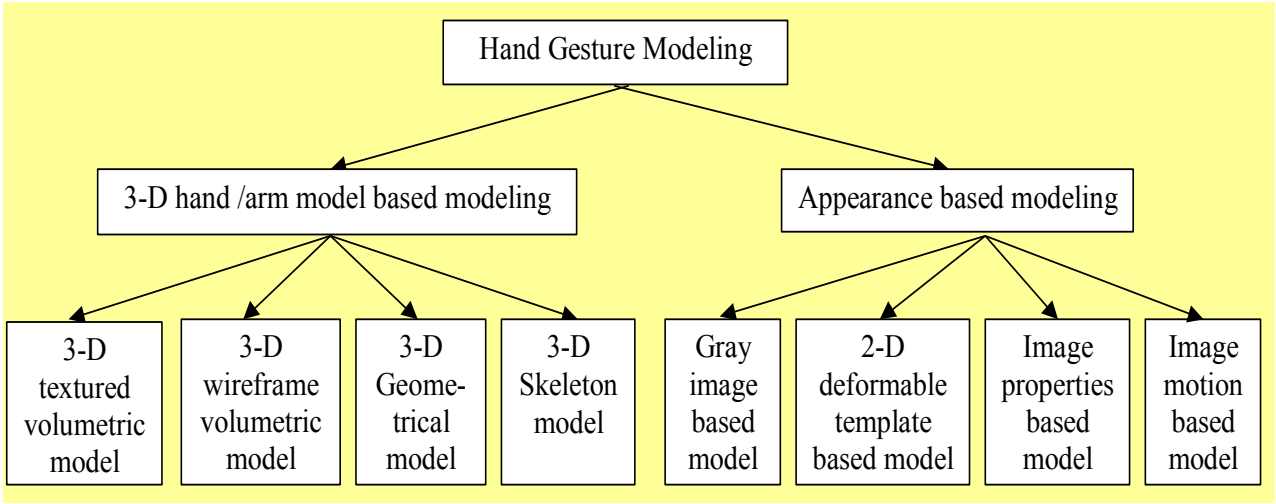
Preprocessing

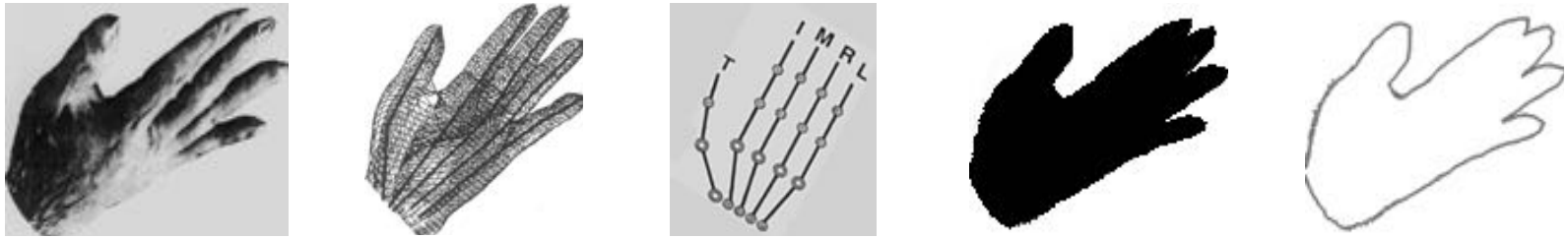
- Pixel level segmentation
  - Background subtraction: Works good on known background or static background
  - Color / texture segmentation
    - ✓ Hand detection
    - ✓ Color marker detection
- Contour detection
  - Not directly depending on skin color and lighting conditions
  - There can be a large number of objects



Preprocessing – Hand Segmentation

- Classification of hand models

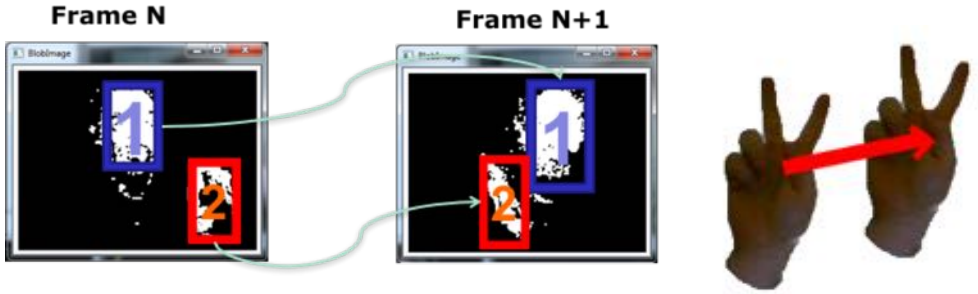




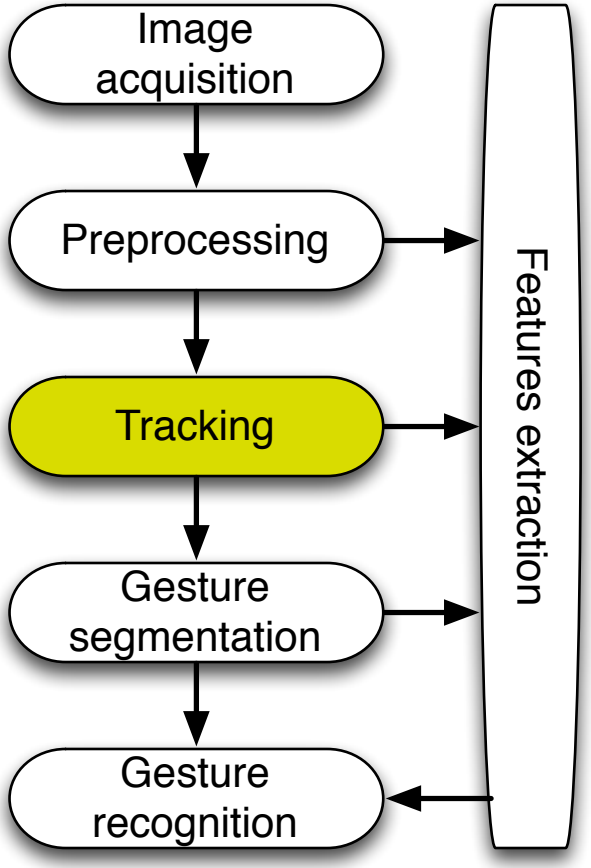
(a) (b) (c) (d) (e)

Fig.3: Representing the same hand posture by different hand models. (a) 3-D textured volumetric model; (b) 3-D wireframe volumetric model; (c) 3-D skeletal model; (d) Binary silhouette; (e) Contour model.

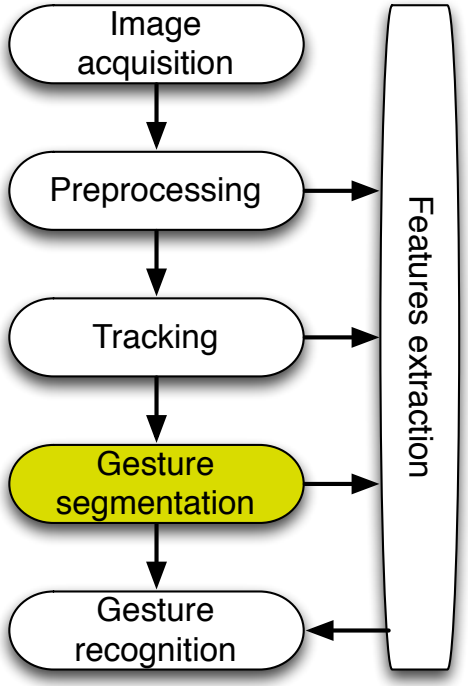
Tracking



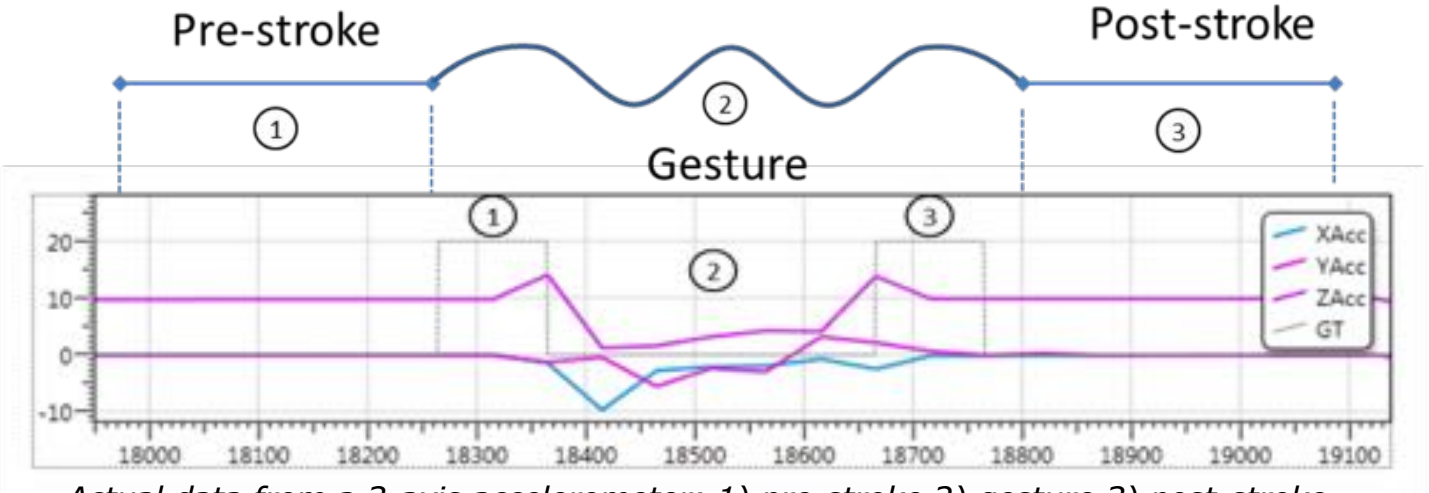
- Sometimes missing frames or occlusion
- Solutions
  - Kalman filter
    - ✓ Estimate current state based on previous
    - ✓ Easily computable in real-time
  - Condensation
    - ✓ Detect and track contours of moving objects in a cluttered environment
    - ✓ One of the most used technique for tracking
  - CAMshift
    - ✓ Fast, real-time
    - ✓ It may be possible to improve accuracy by using different colour representation



Gesture Segmentation



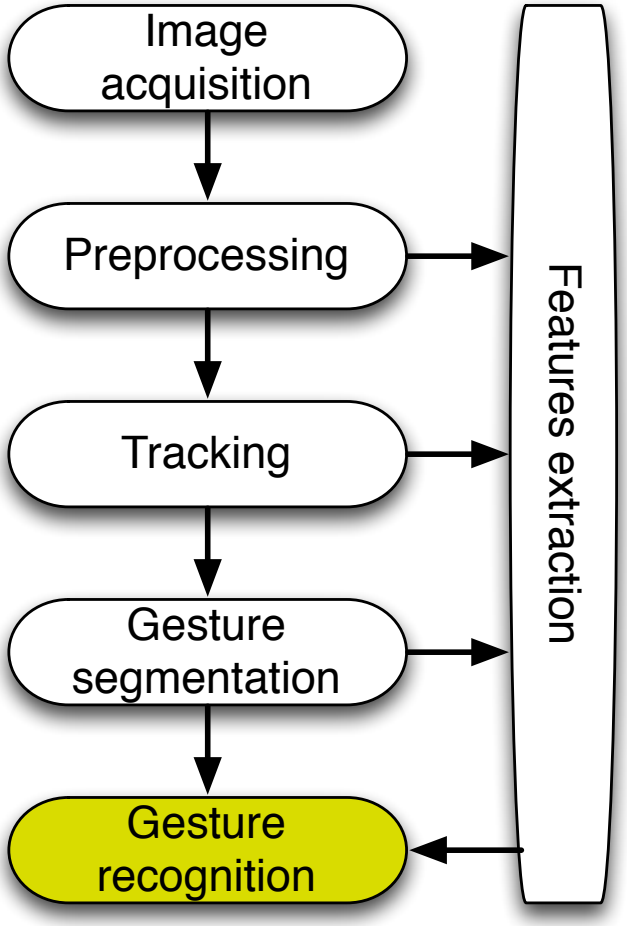
- Gesture decomposition
  - Preparation, stroke and retraction
  - Only the stroke has the information
  - Preparation and retraction prepare and end the stroke
  - Hands are not moving -> end of gesture
- Statistical approach (e.g. HMM)
- Declarative rules



Actual data from a 3-axis accelerometer: 1) pre-stroke 2) gesture 3) post-stroke

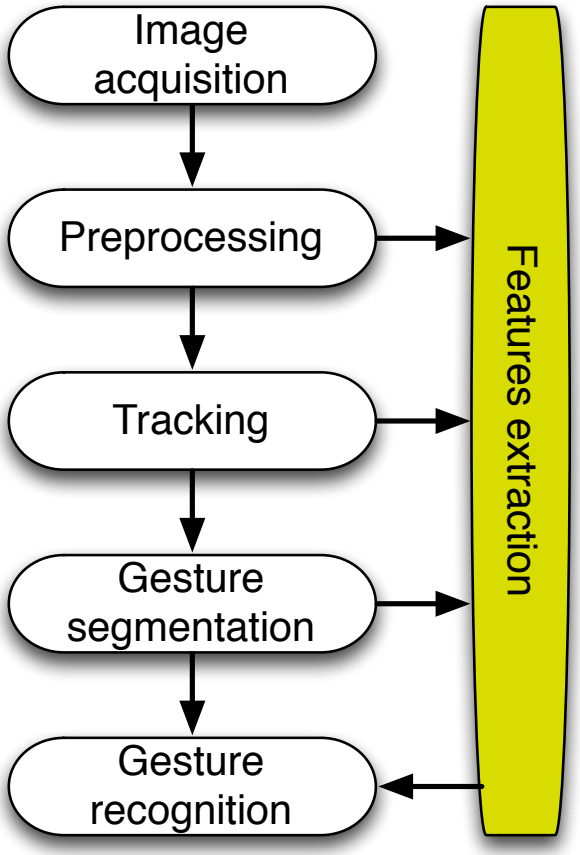
Gesture Recognition

- Three broad families of algorithms
  - Template matching
    - ✓ Linear classifier (e.g. Rubine)
    - ✓ K-means
  - Machine Learning based algorithms
    - ✓ Hidden Markov models
    - ✓ Neural networks
    - ✓ Support Vector Machine (SVN)
  - Rule based approaches
  - Some approaches seek to mix these families and keep the strengths of each



Features extraction

- List of potential features
  - Position, acceleration, velocity
  - Spatial – temporal width
  - FFT of the position
  - Etc.
- The features can be extracted in three steps of the process chain:
  - Preprocessing
  - Tracking
  - Gesture segmentation

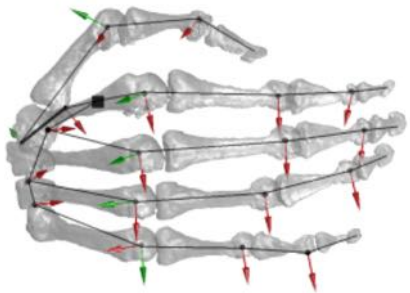


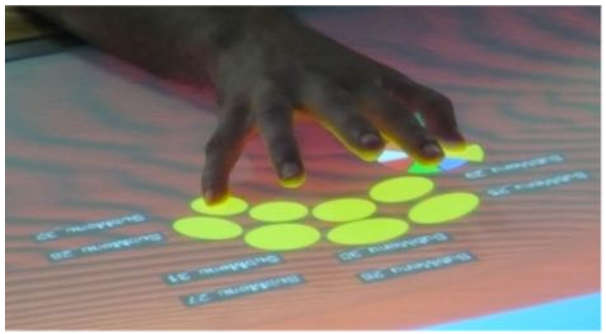
Challenges, pros and cons

- Design challenges
  - Context and lighting conditions (optic systems)
  - Feedback for the user
  - Gesture vocabulary (small – large, type of gesture used, etc.)
  - Real-time interaction
  - Mobile gesture interfaces
  - Multimodality
- Advantages:
  - "Space" effective interaction
  - No need for device
- Drawbacks:
  - "Gorilla arm"
  - User dependent gestures – few universal understandable gestures
  - Computationally expensive

Gestures Vocabulary

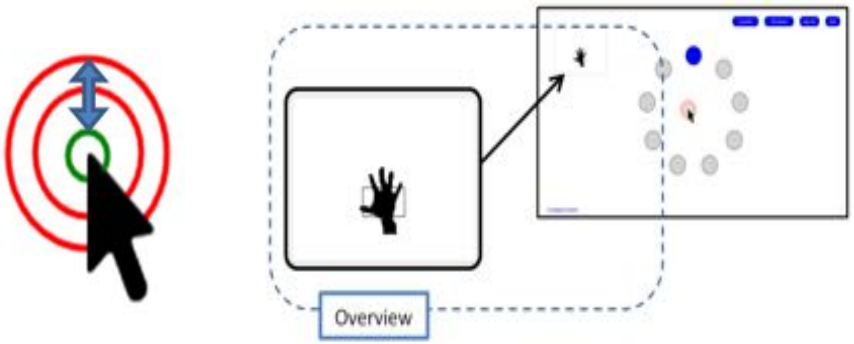
- Choosing a good gesture vocabulary is no easy task!
- Pitfalls:
  - Hard to remember gestures
  - Hard to perform gestures (effort)
  - “Gorilla arm”
- The human body has degrees of freedom and limitations that have to be taken into account… and can be exploited!
- Look at what exists already, what people know
  - Example with the American Sign Language (ASL)
  - Gestures for the alphabet and for the representation of concepts and objects

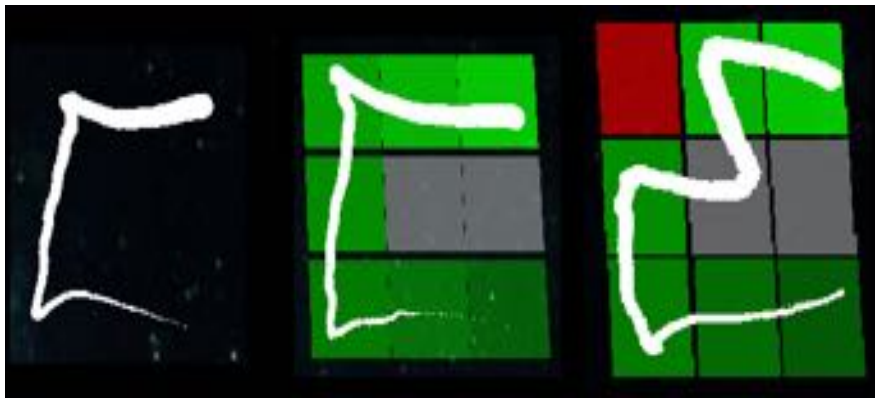


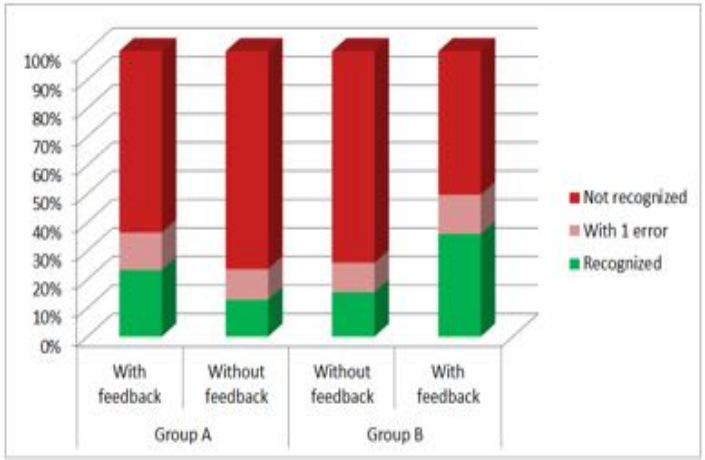


Feedback, feedback, feedback

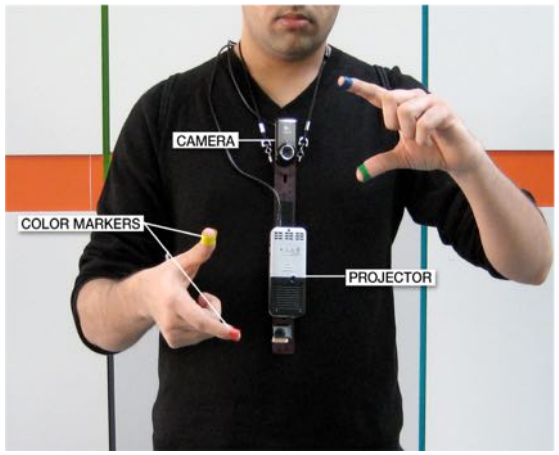
- User perform better with feedback: pointing and other gestures
- Feedback have to be adapted!
  - Can be distracting
- User prefer feedback, especially for learning







Technology

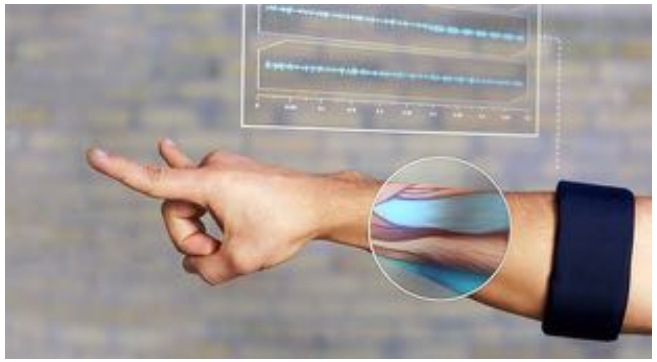


Sixthsense

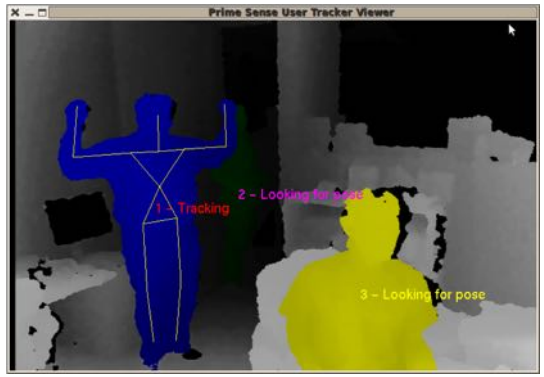
G-speak

Leap

PS Move, Kinect



MYO



Kinect

Tabletop Interfaces

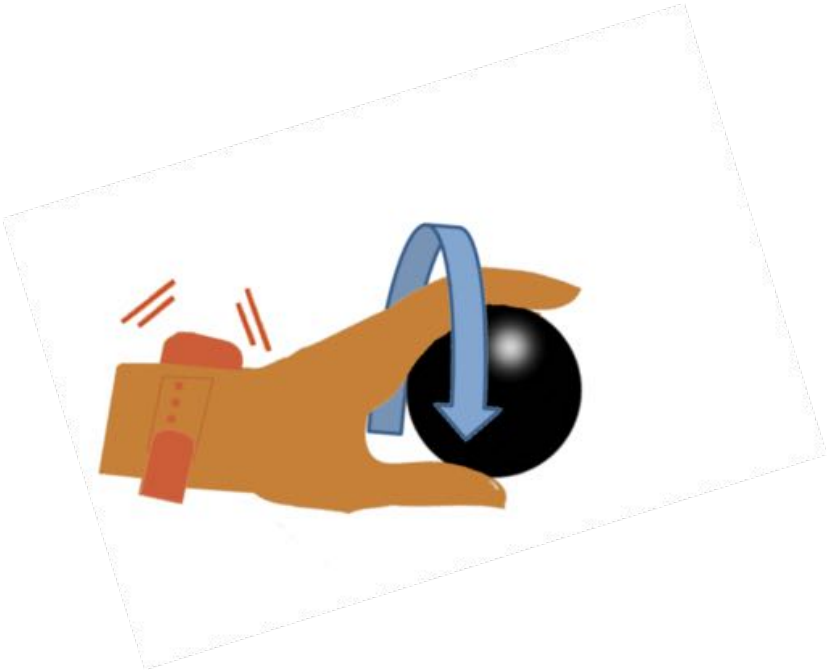
- Tabletop : 2D gestures
- Technologies
  - Video (oldest 2000 or before)
  - Marker-based tables
    - ✓ Example: ReacTiVision (2007)
  - Capacitive coupling
    - ✓ Example: Circle Twelve's DiamondTouch (2001)
  - FTIR (Frustrated Total Internal Reflection)
    - ✓ Example: Jeff Han's multi touch surface (2006)
  - Optical glass
    - ✓ Example: Microsoft's PixelSense (2011)

Tabletop Interfaces

- Tabletop interfaces are also gesture interaction devices, but with a number of very specific features:
  - Continuous input as soon as the user touches the surface
  - No gesture segmentation problem
  - Multi-finger + multi-user = LOTS of input events to manage
  - Some surfaces can also recognise objects
- Multi-Touch: hundreds of input events per second
  - ✓ User identity?
  - ✓ Just use normal GUI?
  - ✓ Complex gesture involving multiple fingers?
- Multi-User
  - there is enough room for a few people to sit around!
  - Not everybody sees the same thing
  - Studies have shown that users adopt territories on a tabletop, to help coordination

Gestures: Some Open Questions
(with no clear answer)

- Do we really want big, full body gestures at all? vs. small, “lazy” gestures
- The hand as a pointer: is it right to consider the hand as yet another pointing device? In which case?



What You Should Be Able to Answer

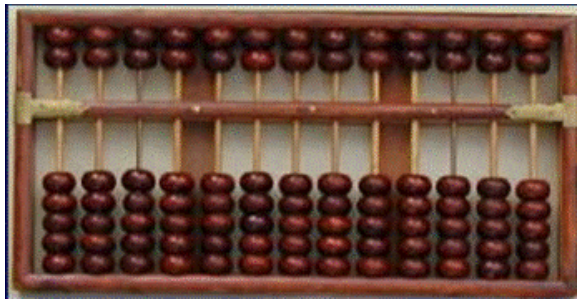
- What is voice-based interaction?
- How does work voice recognition and voice synthesis?
- Explain challenges of voice-based interaction system

- Give a definition of a gesture
- Describe a pipeline for gesture recognition
- Describe how to define a good gesture vocabulary
- Give examples of tabletop technologies

[4.3]Tangible Interfaces

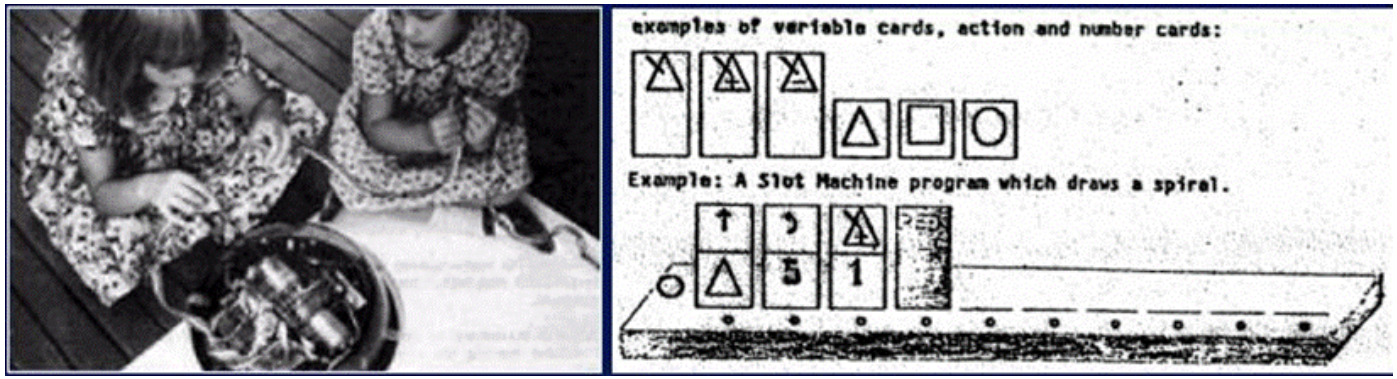
Historical~8000 BC:

- Long history, diversed practices:
  - Communication, education, reasoning, self-representation, spirituality
  - Clay Accounting Tokens (Mesopotamia)
  - ~3000 BC: Board games (Mesopotamia)
  - 1200 AD..today: Abacus (Japan, China)
- Physical games / physical calculator
  - Externalize cognition
  - Manipulation power
  - What you see is what I see (collaborative)
  - We manipulate the same objects (collaboration)
- Physical representation
  - abstract operations / rules
  - abstract numerical values
- Physical control

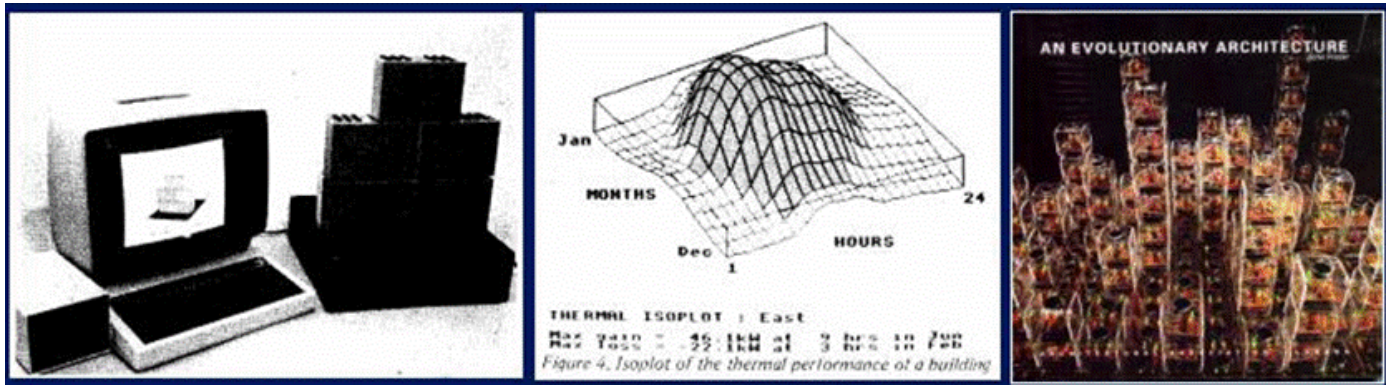


Historical context

- 1976: Perlman, Hillis (MIT), Education domain
  - Tools for children to physically experiment with programming concepts such as recursion

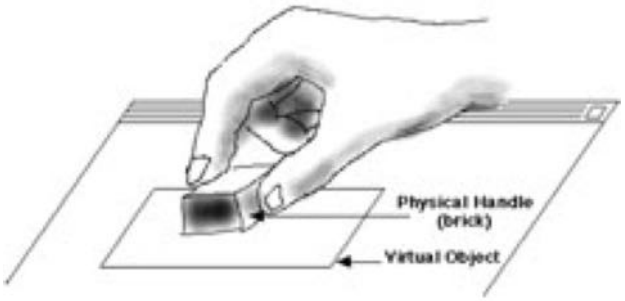


- 1980: Aish, Frazer (UK) Architectural domain
  - Building blocks, Tools to facilitate communication



TUI History

- *Pierre Wellner, Wendy Mackay, and Rich Gold. 1993. Back to the real world. Commun. ACM 36, 7 (July 1993), 24-26.*
- Graspable User Interface (1995)
  - Physicality: Realizing the underexploited potential of hands as interfaces
- Mix of virtual and physical artifacts
- Artifacts have clear affordances
  - i.e., the actions they allow are clearly derivable from aspect and constraints
- Direct control of virtual objects through physical handles and tools
- Foundation of tangible interaction

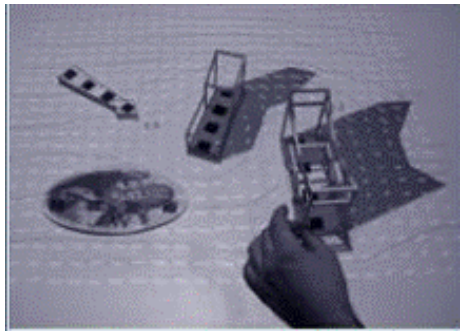
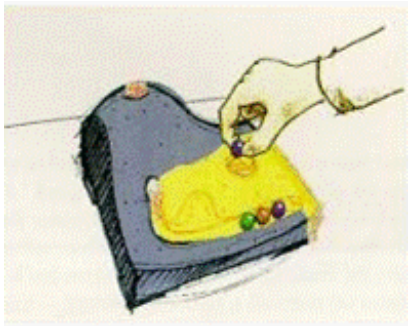


Fitzmaurice, George W., Hiroshi Ishii, and William AS Buxton. "Bricks: laying the foundations for graspable user interfaces." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 442-449. ACM Press/Addison-Wesley Publishing Co., 1995.

Tangible Interaction: A First Outlook

- Computation that moves beyond the desktop
- Giving physical form to digital information
  + interaction capabilities
- Tangible interfaces are at the border between the physical and the digital
  - ➢We live on the border where bits meet atoms. In the flood of pixels from the ubiquitous GUI screens, we are losing our sense of body and places. [Ishii, 1997]

- "*Direct link between the digital bit and the physical atom*"


Marble Answering Machine (1992)

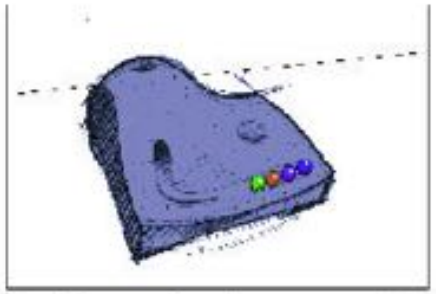
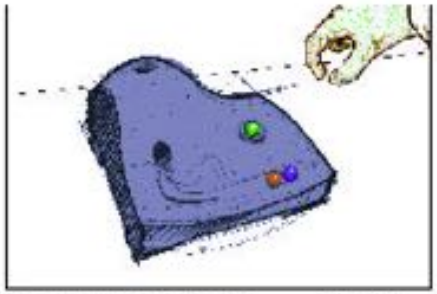


Figure 6.i Incoming messages await...



The user listens to a message... Figure 6.ii

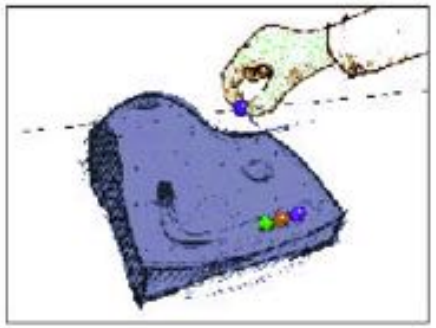
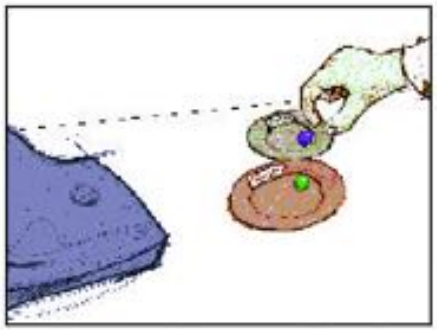


Figure 6.iii ...the user moves the message

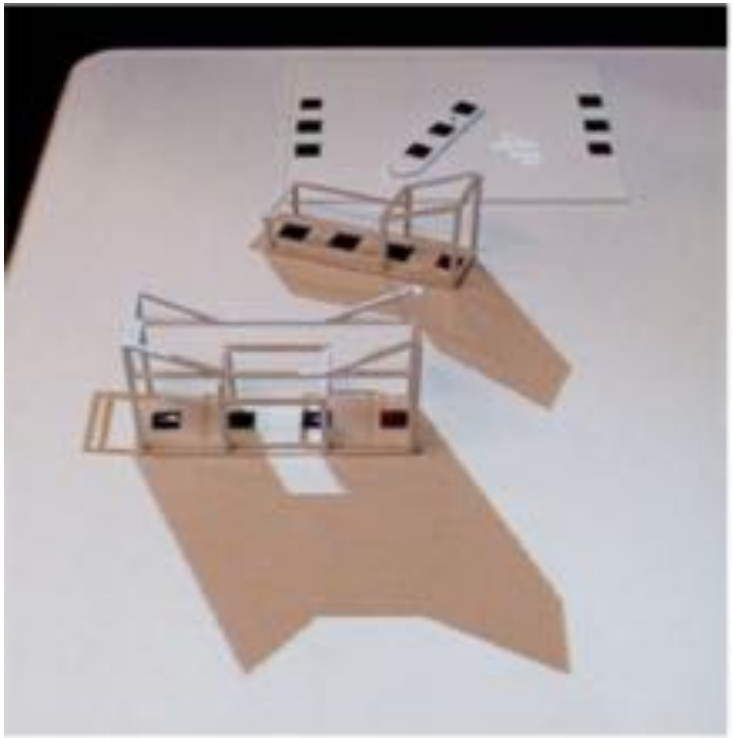


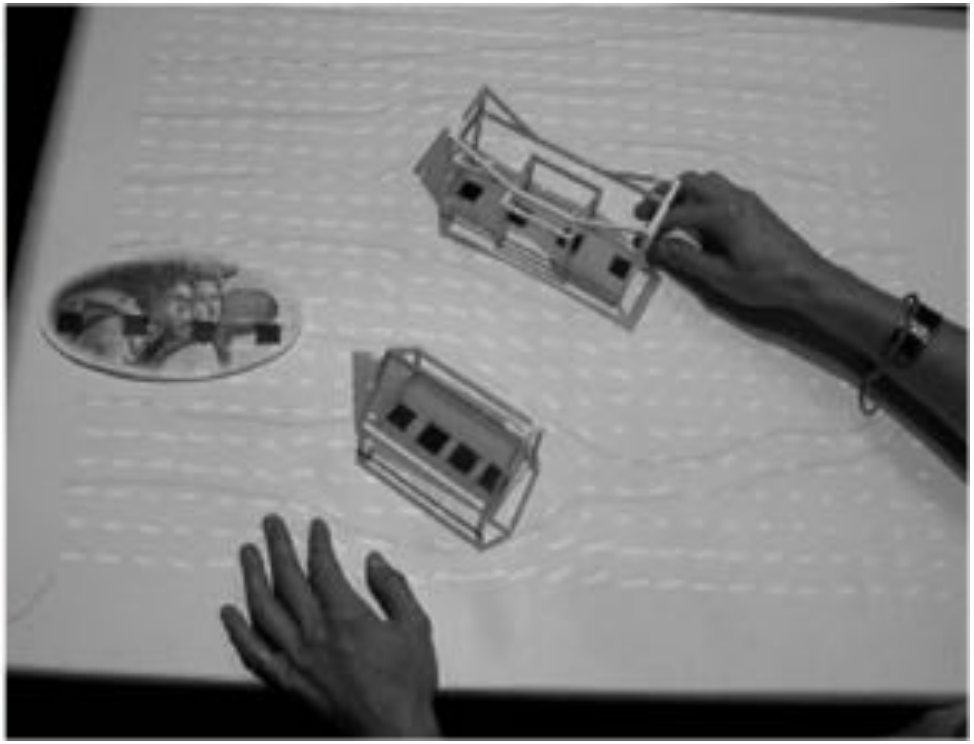
...to each roommate's in-tray. Figure 6.iv

MusicBottles (1999)

- Affordance?

URP (1999)





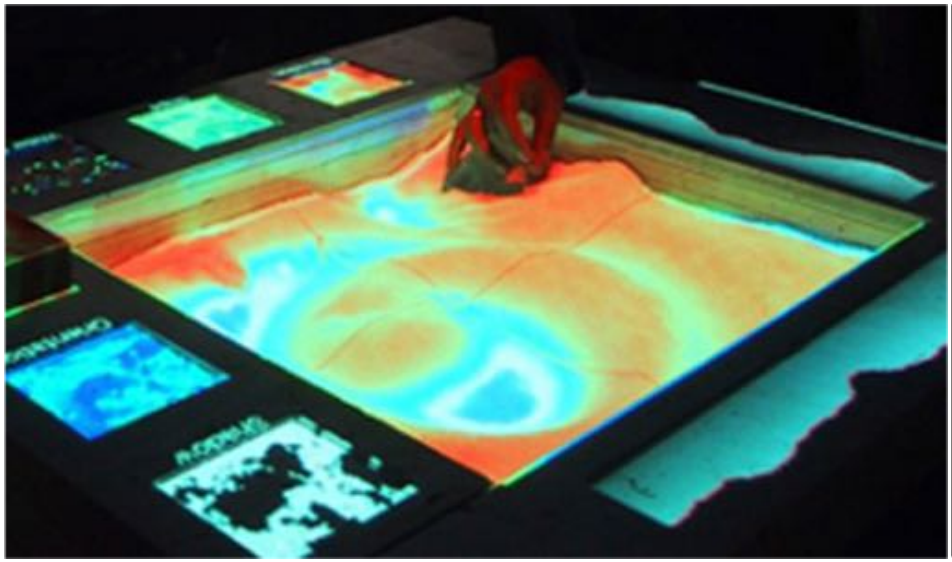
- Digital representation linked to a real context

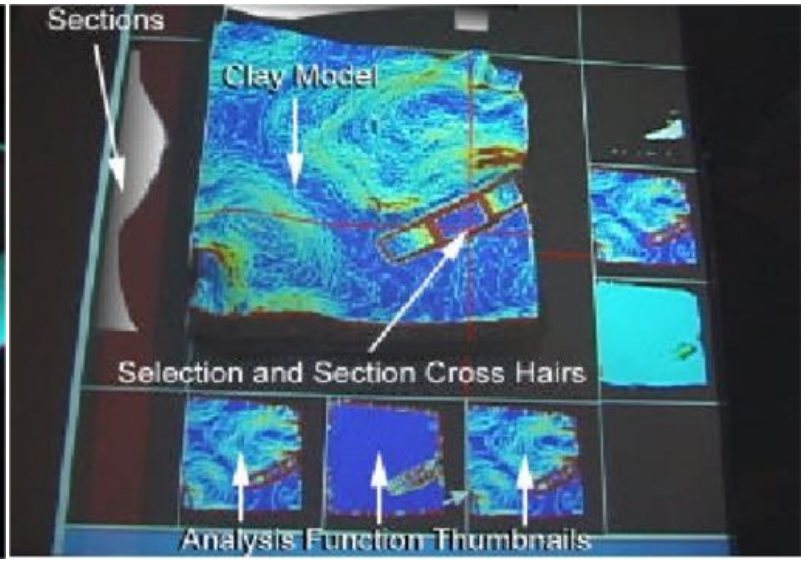
Audiopad & Reactable

- Multi-user interaction and collaboration

Curlybot (1999), topobo (2007)

Sandscape & Illuminating Clay





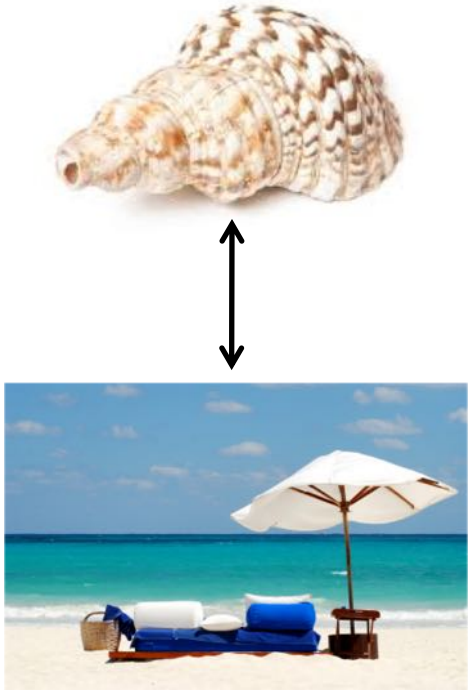
- Sand and clay as interface

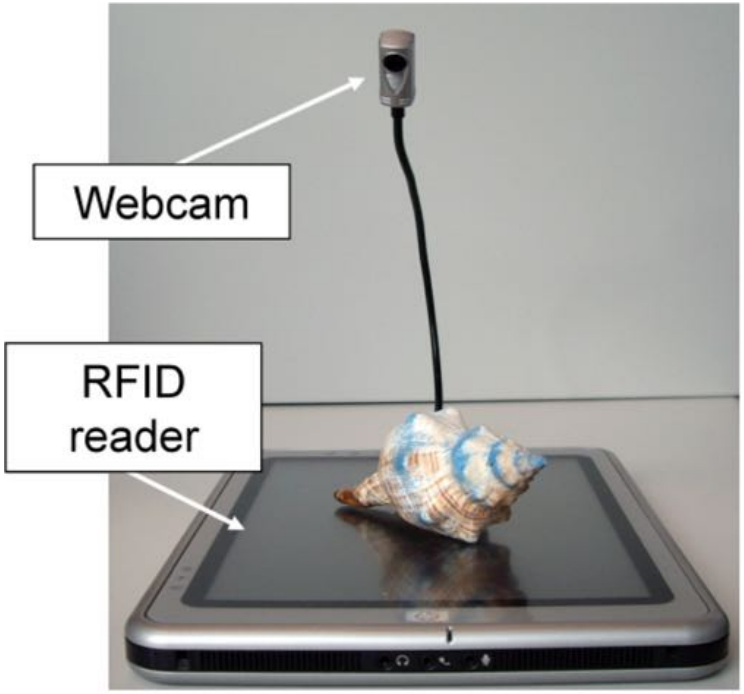
I/O Brush (2005)

- Virtual or physical tool?

Memodules

- Associating personal objects with digital artefacts
  - Example: that shell you picked up during your holidays could be linked to your holiday pictures and movies





inTouch (1998)

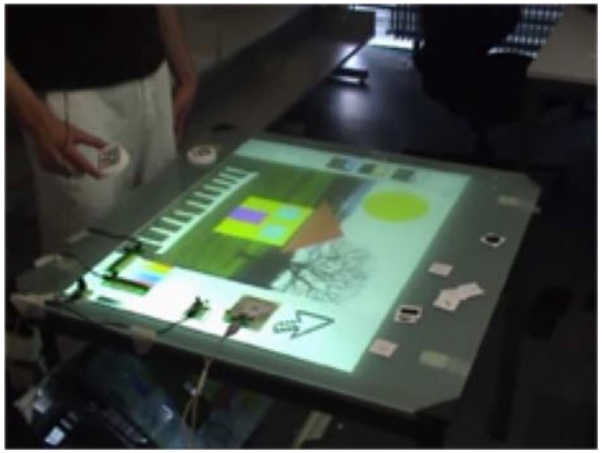
http://vimeo.com/44537894

Definition and characteristics

- Tangible Interaction: umbrella term
  - Tangible user interfaces
  - Graspable user interfaces
  - Embodied interaction
- Characteristics
  - Physical Artifacts that give physical form to digital information
  - Serves simultaneously as:
    - ✓ input and output
    - ✓ control and representation
- Advantages
  - Physical embodiment of data
  - Direct manipulation / Spatial thinking
  - Multi-user interactions / communication
  - Users' interaction in real contexts
  - Persistency
  - Strong affordances

Multimodal vs. Tangible Interaction

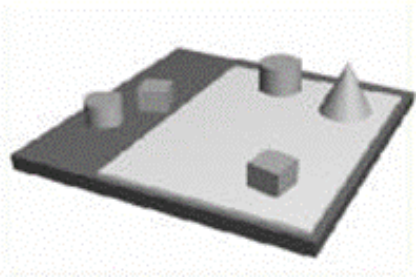
- In a narrow sense, tangible interaction can be considered as the use of one (or a subset of) modalities
- Multimodality suggest:
  - New physical and social contexts for interaction
  - New mediums of engagement with computation
- Divergence from traditional GUI interaction paradigm
  - Progressive disembodiment of interface (e.g. speech, gesture, etc.)
  - Progressive embodiment of interface (e.g., tangible interfaces)



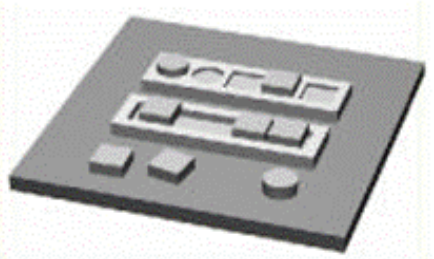
Xpaint
(MMI 2006)

Survey of illustrative systems:
Common structural approaches

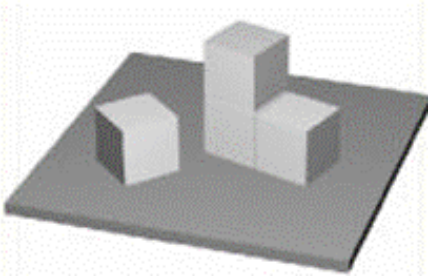
1. Interactive surfaces



2. Tokens and constraints

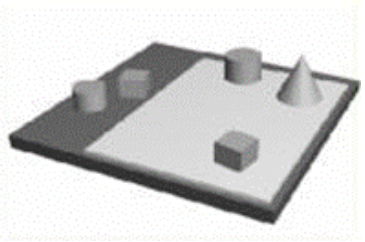


3. Constructive assemblies



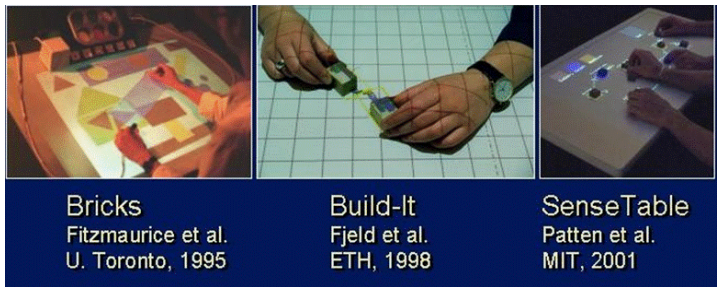
Ullmer, CUSO winter school Multimodal interfaces 2004

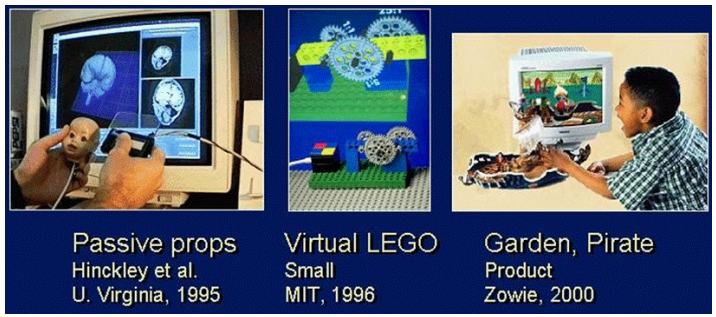
1. Interactive surfaces



Use of surfaces as interactive workspaces for manipulating physical tokens

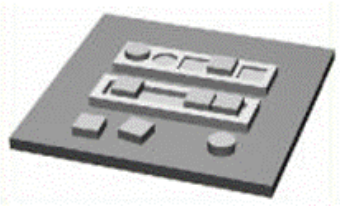
- **Graspable handles**: dynamically-rebindable « handles » for physically manipulating graphical content
- **Spatially manipulable models**: permanently bound « physical icons » for manipulating simulations
- **Magic mirrors**: Use of physically tracked objects in front of an augmenting display
- **Interactive walls** – postIts, brainstorming, communication (clearBoard, Ishii 1995)



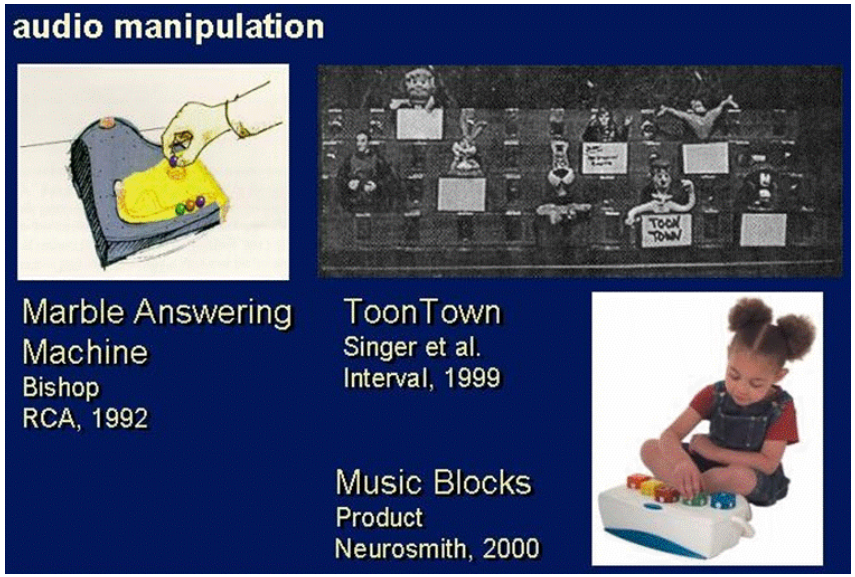


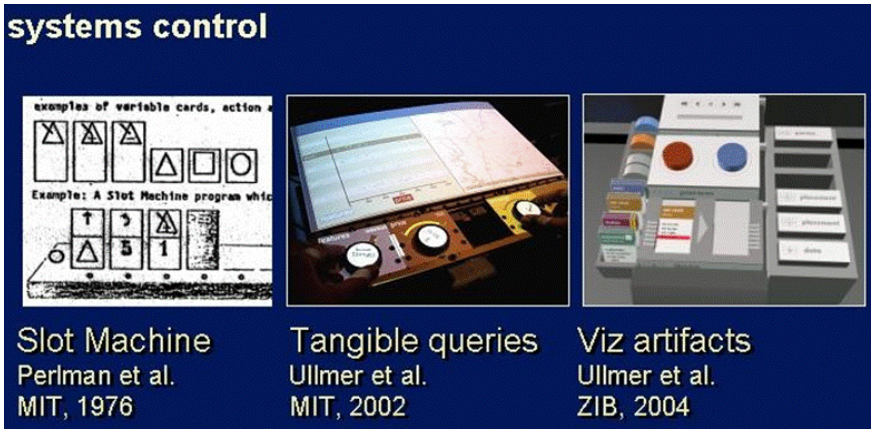
- Pro: builds upon GUI; strong colocated use of graphical mediation
- Con: space consuming; weaker physicality

2. Tokens and constraints



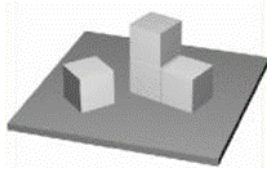
- Audio manipulation
- Image and video manipulation
- System control



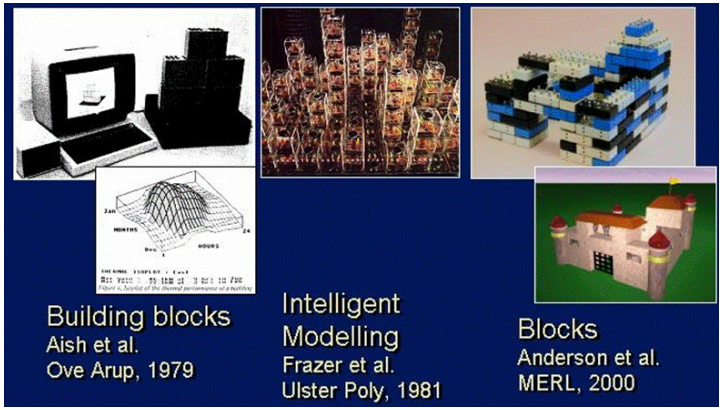


- ➢ Pro: strong use of physicality; complements other approaches
- ➢ Con: less malleable than surfaces

3. Constructive assemblies

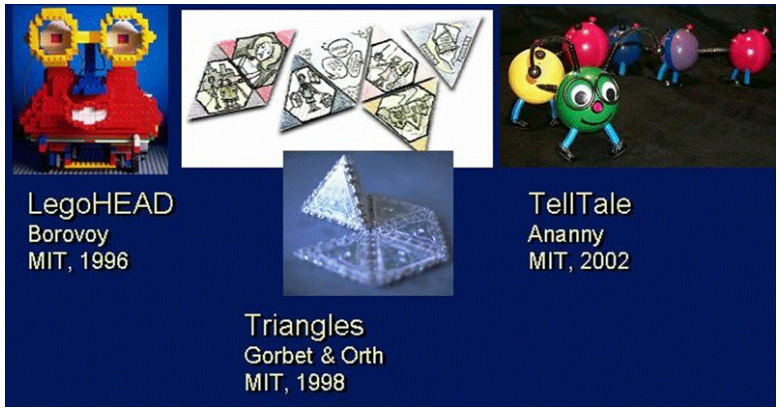


Spatial assembly

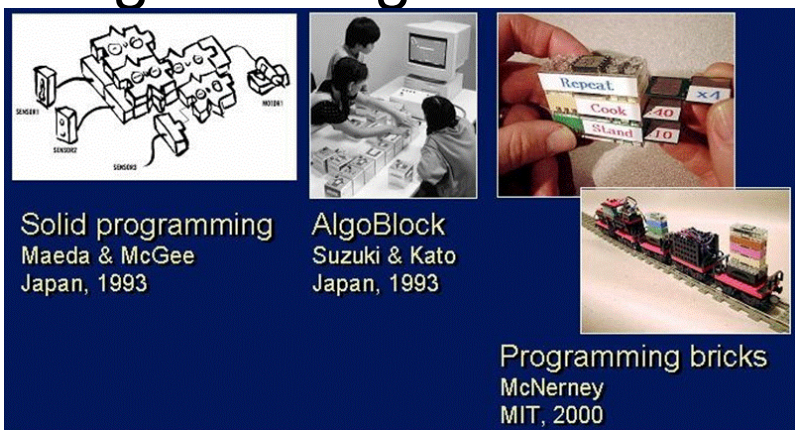


Cellular automata

Storytelling



Programming - education

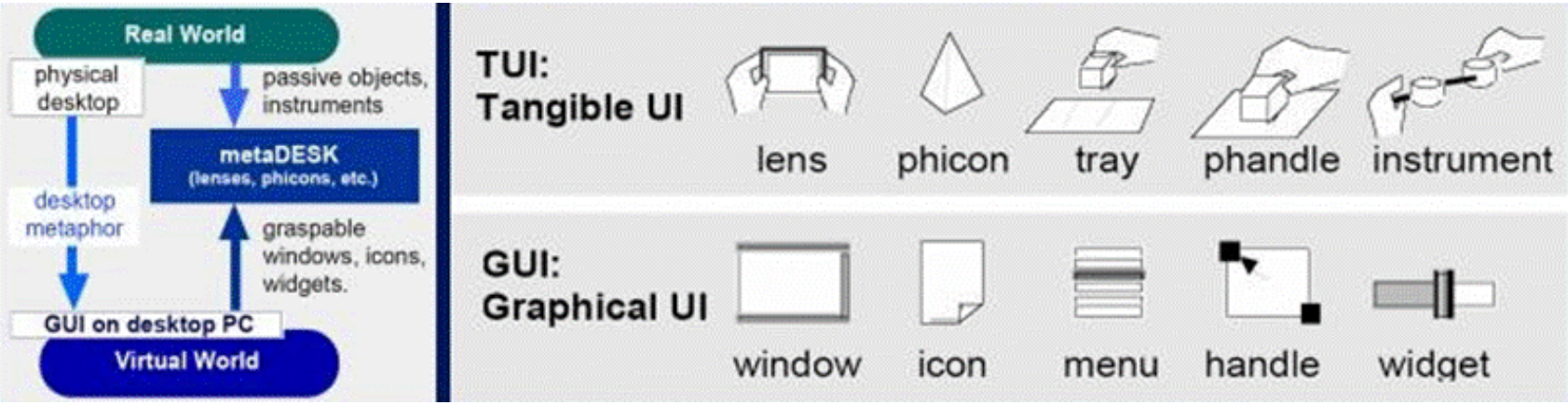


- Pro: strong tradition, expression of physicality
- Con: scalability, abstraction have seemed difficult

Taxonomies of TUIs

- Tangible interfaces cover many different types of interfaces, with the tangible aspect as core property
- Multiple classifications and taxonomies have been proposed, which allow to understand the design space
  - Ullmer/Ishii, 1997
  - Ullmer/Ishii, 1999
  - Holmquist et al., 1998
  - Ullmer/Ishii 2004-05, Shaer et al. 2003
  - Van den Hoven/Eggen, 2004
  - Fishkin, 2004

Theory for TUIs
GUI/TUI analogs (Ishii & Ullmer, 1997)

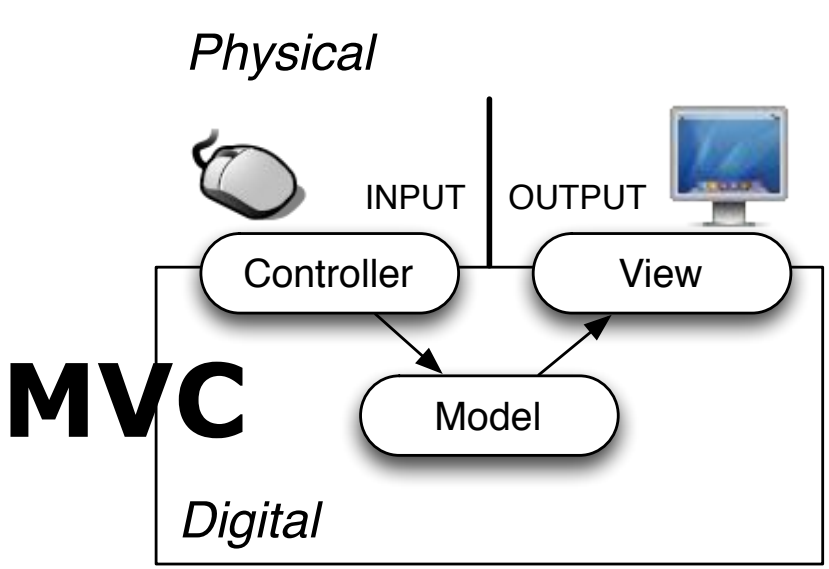


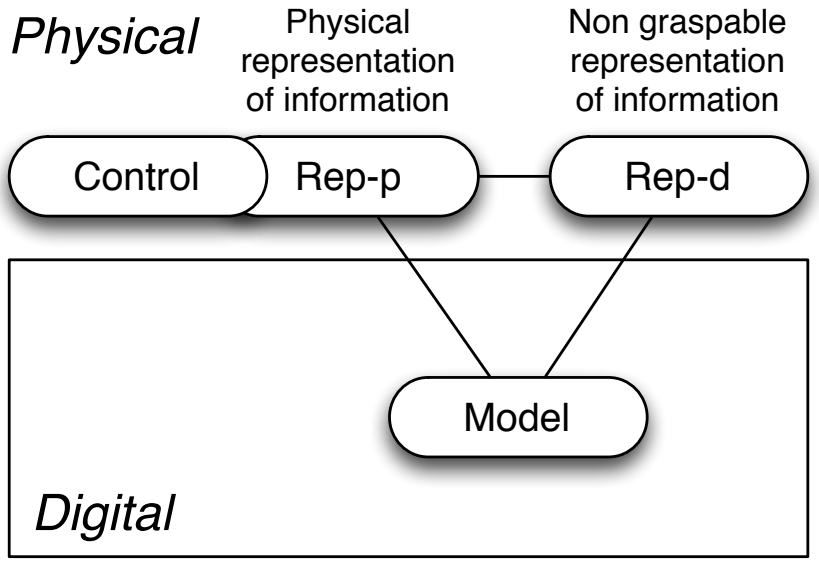
- Introduced terminology that become common;
  - E.g., « physical icon » = « phicon »
- Pro: builds upon huge experience base with graphical interfaces
- Con: there are fundamental differences between graphical and tangible interfaces; analog risks danger of conceptual & counter-productive research trajectory

Ishii, Hiroshi, and Brygg Ullmer. "Tangible bits: towards seamless interfaces between people, bits and atoms." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 234-241. ACM, 1997.

Interaction Model: the MCRpd (1999)

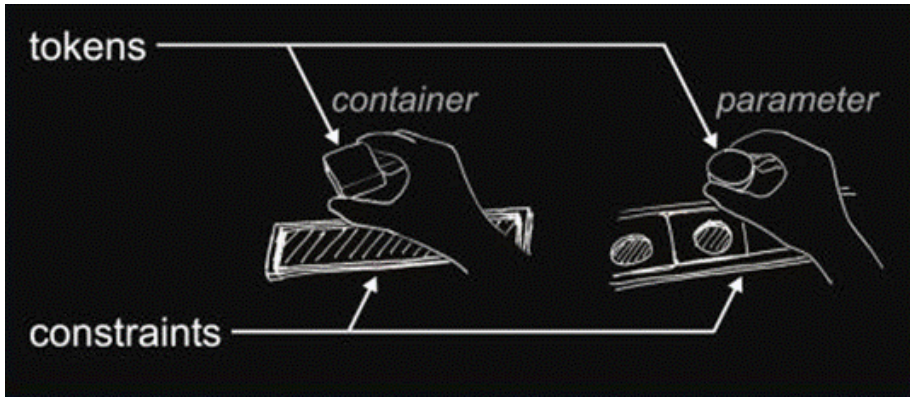
- MCRpd (Model – Control – Representation: physical & digital)
- Separates the View in two components:
  - *Physical representation* (rep-p): artifacts constituting the physically embodied components of tangible interfaces
  - *Digital representation* (rep-d): computationally mediated components of tangible interfaces (no embodied form)

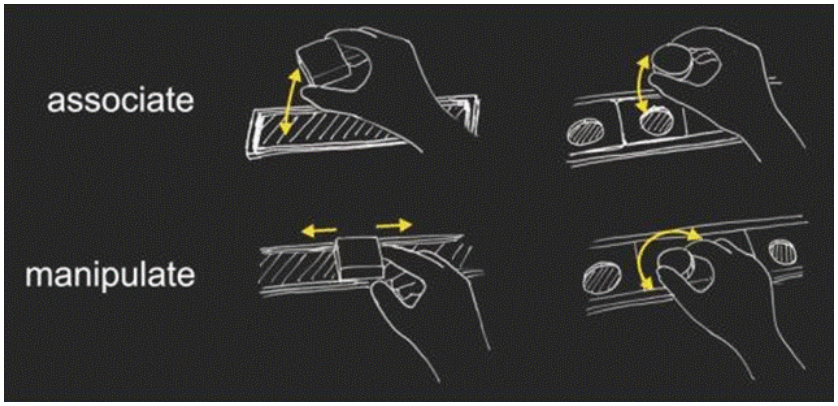




Ullmer, Brygg, and Hiroshi Ishii. "Emerging frameworks for tangible user interfaces." IBM systems journal 39.3.4 (2000): 915-931.

TAC: Token + constraint characterization (Shaer 2004)



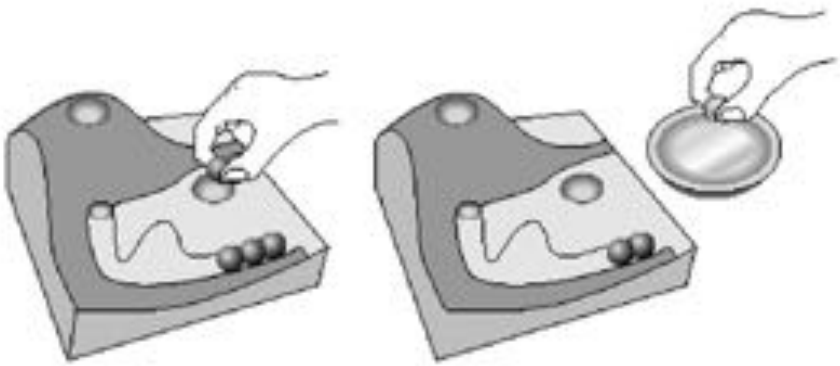


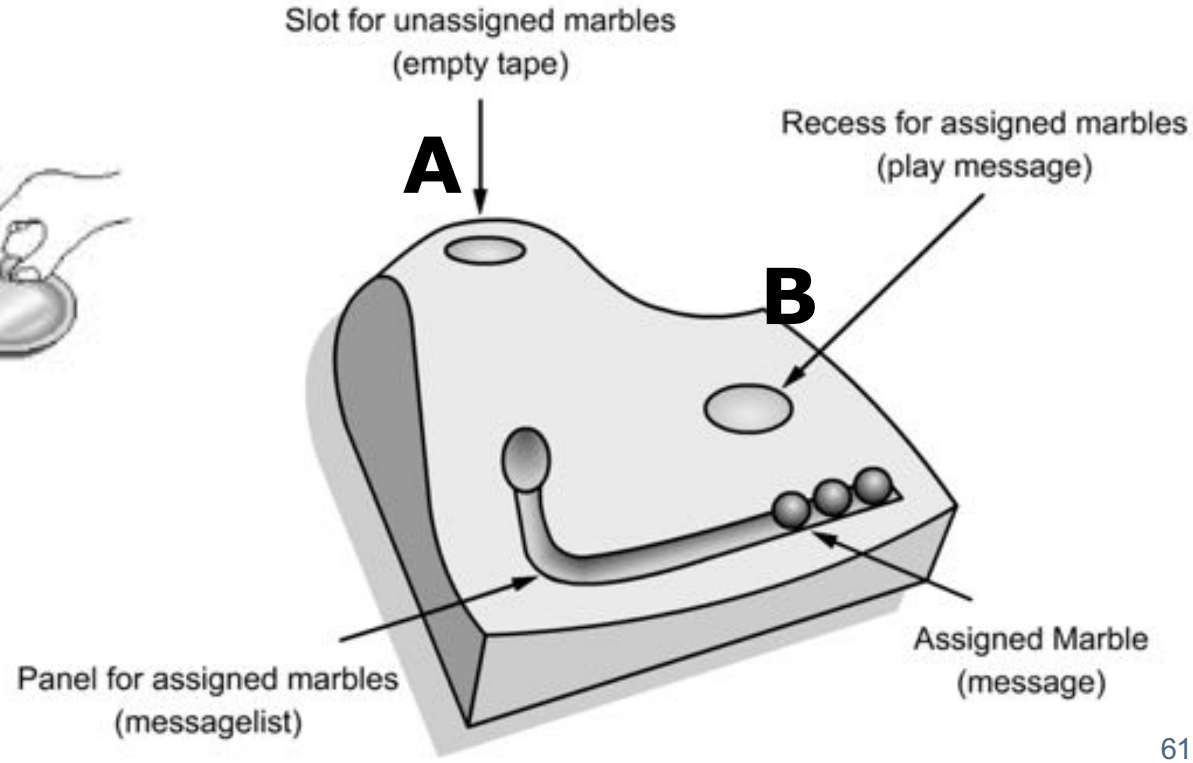
- Highlights semantic role of constraints / reference frames in which tokens are manipulated « grammatical » constructions
- Pro: exposes semantics for different stages of interaction
- Con: limited scope of interaction

Shaer, Orit, et al. "The TAC paradigm: specifying tangible user interfaces." Personal and Ubiquitous Computing 8.5 (2004): 359-369.

TAC example

TAC	Association			Behaviour	
	Token	Constraints	Variable	Action	Feedback
1	Marble	Slot A	Message	Remove	Marble is removed
2	Marble	Slot B	Message	Add	Message is played
				Remove	Message is stopped





Taxonomies of TUIs: Fishkin

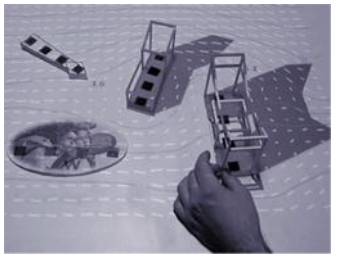
- Taxonomy along two dimensions:
  - Metaphor (Noun ~ shape of the object ; Verb ~ motion of the object)
  - Embodiment (how closely tied are I / O)
- "Spectrum" of how much an interface is (or is not) tangible

Metaphor / Embodiment	None	Noun	Verb	Noun and Verb	Full
Full					
Nearby					
Environ-ment					
Distant					

Fishkin, Kenneth P. "A taxonomy for and analysis of tangible interfaces." Personal and Ubiquitous Computing 8, no. 5 (2004): 347-358.

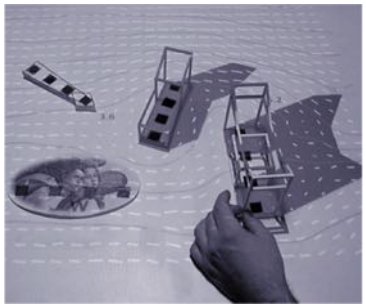
Taxonomies of TUIs: Fishkin (cont.)

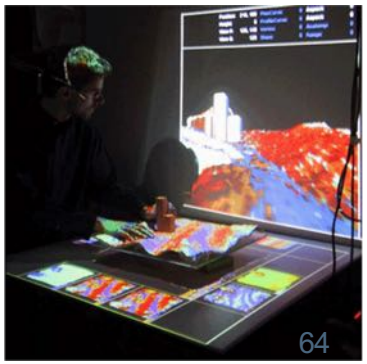
- Full embodiment
  - The output is the input device
  - Input output coincidence (ex.: Curlybot)
- Nearby embodiment
  - The output is tightly coupled to the focus of the input (ex.: URP)
- Environmental embodiment
  - The output is “around” the user
  - (typically, audio output) (ex.: ToonTown)
- Distant embodiment
  - The output is “over there” on another screen, or even another room, like a remote control (ex.: Doll’s Head)



Taxonomies of TUIs: Fishkin (cont.)

- Noun metaphor
  - Object looks like the real thing (ex.: tagged objects)
  - However, actions employed on/with that object are either not analogous or only weakly
- Verb metaphor
  - Object acts like the real thing (ex.: Bricks' "pens")
- Noun & verb metaphor
  - Object looks and acts like the real thing
  - … but they are still different (ex.: building in URP)
- Full metaphor
  - The virtual system is the physical system
  - The users need make no analogy at all
  - "Really direct manipulation" (ex.: Illuminating Clay)





Implementation technologies

- RFID – Radio Frequency IDentification
  - Wireless radio-based technologies
  - Passive vs active tags
  - Mono vs multiple tags reader
- Microcontrollers, sensors and actuators
  - Microcontroller: small computers embedded in a physical object
  - Sensors (accelerometers, temperature, …) and actuators (led, motors,…)
  - Arduinos, phidgets
- Computer vision
  - “Visual Tags”
  - light-weight projector for providing real-time graphical output
  - computer vision software package

Some GUIs / TUIs differences

Graphical User Interface (GUI)	Tangible User Interface (TUI)
Separates input from output	Combines input and output
Separates control from representation	Integrates control and representation
Enforces visual interaction	Supports visual, but also spatial and relational interaction
Icons (digital)	Phicons (physical)
Painted bits	Tangible bits

What You Should Be Able to Answer

- What are tangible interfaces?
- Give a definition of tangible interaction
- What are the goals of tangible interaction?
- Understand the difference between
  - the MCRpd interaction model
  - the TAC model
  - the taxonomy of Fishkin (axes)
- Compare GUIs and TUIs and MUIs