Multimodal Interaction

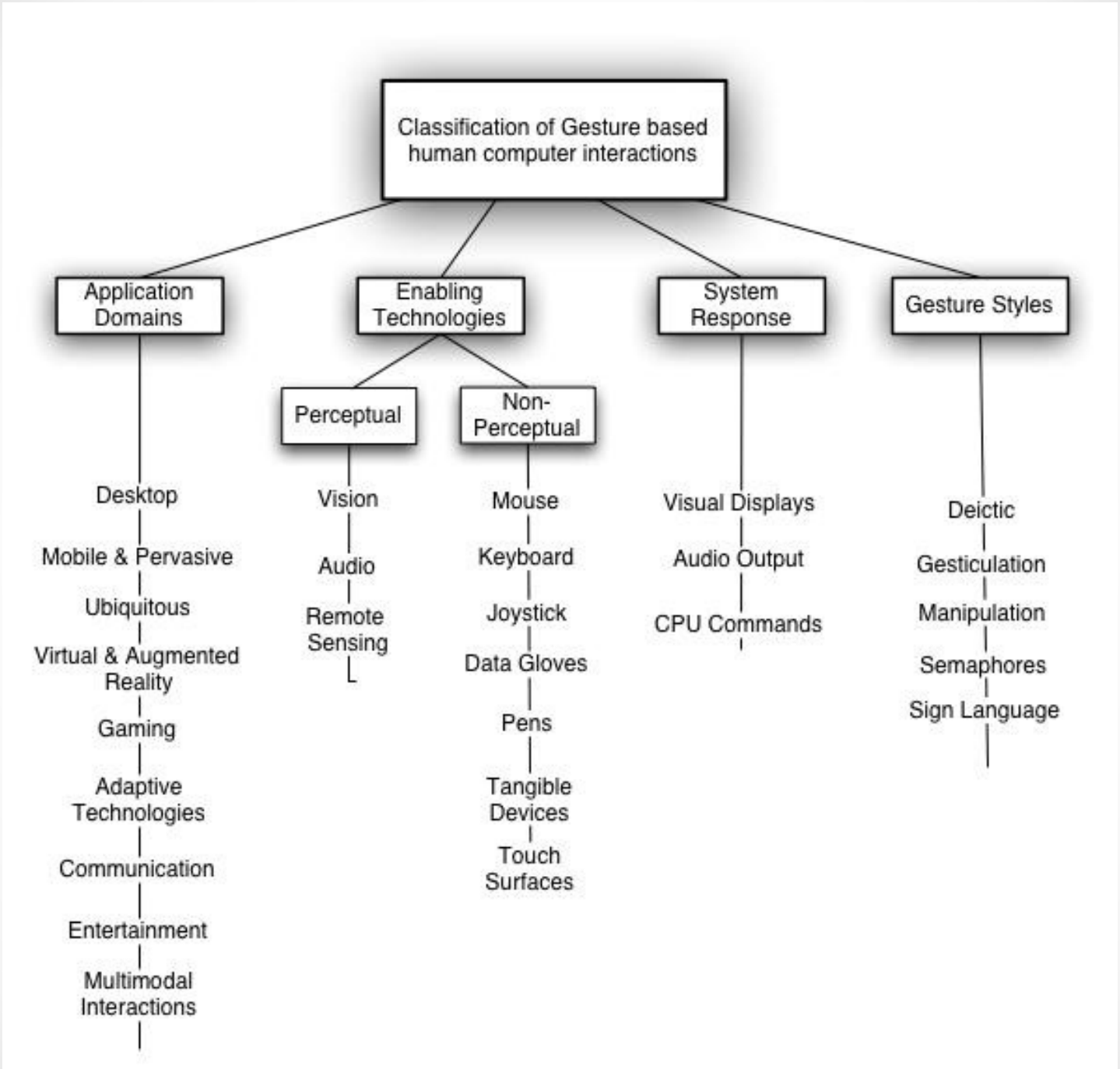
Lesson 5
Gesture Interaction

Maria De Marsico
demarsico@di.uniroma1.it

Gesture Based Interaction Taxonomy

- Almost every form of possible human gesturing can provide **natural** and intuitive ways to interact with computers
- Almost all input and output technology has been used to enable gesture-based interactions.
- A universally accepted taxonomy would provide a unified perspective of gestures within the field of computer science
- Gestures exist in different forms within different application domains.
- Within the domains, Karam and and Schraefel also consider the various I/O devices to create a (possible) taxonomy.
- Four categories :
  - Gesture style,
  - Application domain,
  - Enabling technology (input)
  - System responses (output)

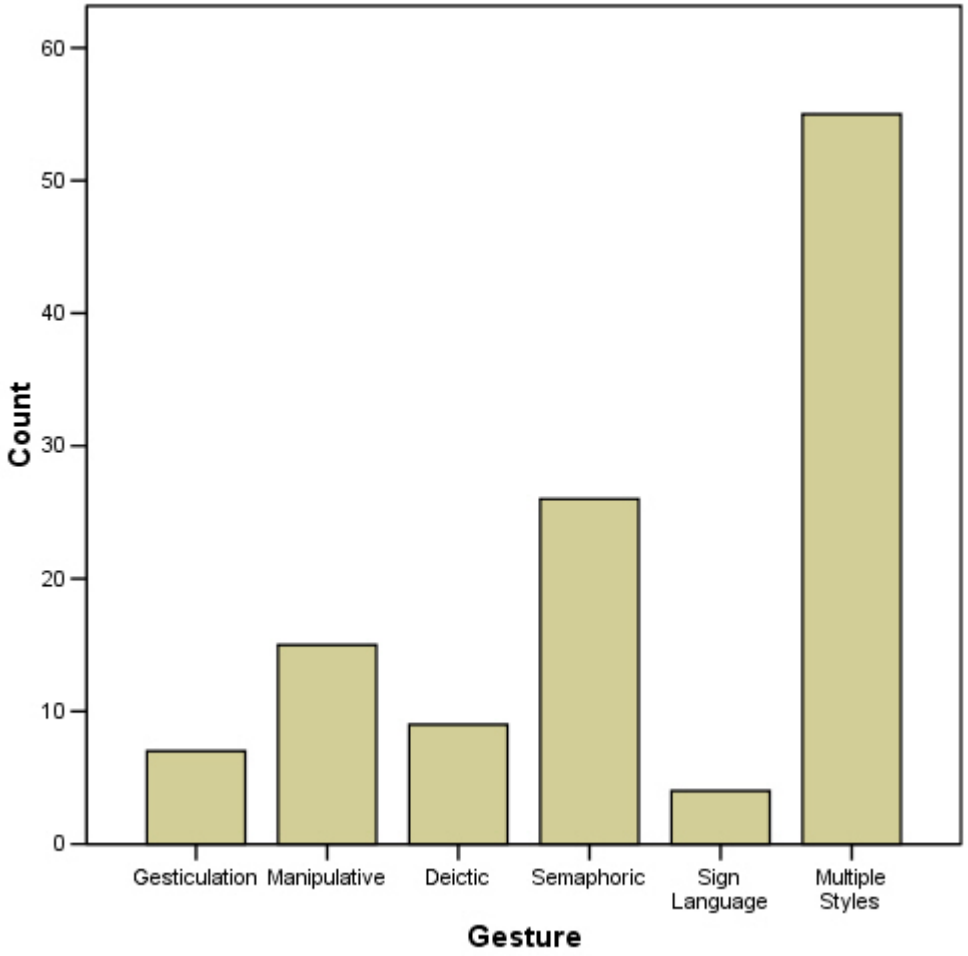
- This lesson will illustrate their taxonomy (see readings)


Classification of Gesture based human computer interactions
Application Domains
Enabling Technologies
System Response
Gesture Styles
Perceptual
Non-Perceptual
Desktop
Mobile & Pervasive
Ubiquitous
Virtual & Augmented Reality
Gaming
Adaptive Technologies
Communication
Entertainment
Multimodal Interactions
Vision
Audio
Remote Sensing
Mouse
Keyboard
Joystick
Data Gloves
Pens
Tangible Devices
Touch Surfaces
Visual Displays
Audio Output
CPU Commands
Deictic
Gesticulation
Manipulation
Semaphores
Sign Language

Gesture Styles

- Deictic or pointing gestures
- Manipulation
- Semaphores
- Gesture-speech = gesticulation, or gesture and speech interfaces where gestures accompany speech for a more 'natural' interaction using bare hands
- Language gestures

Gesture Styles (GS)



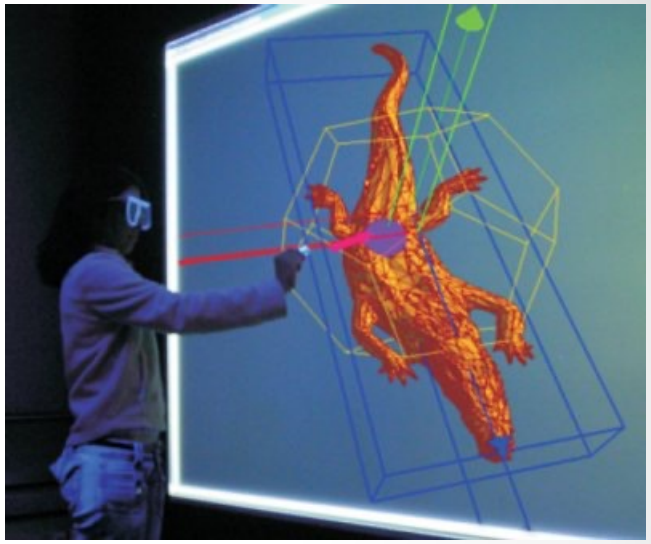
Survey by Karam and Schraefel 2005

GS - Deictic gestures

- Deictic gestures involve pointing to establish the identity or spatial location of an object within the context of the application domain
- **Informative**
- The first example is Bolt's "Put that there" (1980)
  - Deictic gestures are used in conjunction with speech input in an interaction that allows the user to point at a location on a large screen display in order to locate and move objects.
- Deictic gestures are also used
  - to identify objects in virtual reality applications
  - to identify objects to others in CSCW applications
  - for targeting applicances in ubiquitous computing
  - for desktop applications
  - for communication applications

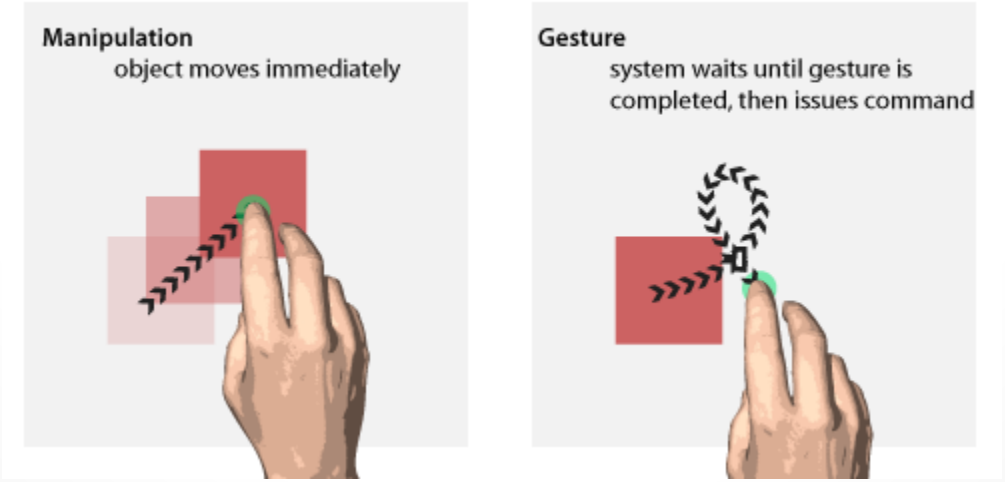
GS - Manipulative gestures

- "A manipulative gesture is one whose intended purpose is to control some entity by applying a tight relationship between the actual movements of the gesturing hand/arm with the entity being manipulated".
- Manipulations can occur
  - on the desktop in a 2-dimensional interaction using a direct manipulation device such as a **mouse** or stylus
  - as a 3-dimensional interaction involving empty handed movements to mimic manipulations of physical objects as in virtual reality interfaces
  - by manipulating actual physical objects that map onto a virtual object in tangible interfaces.



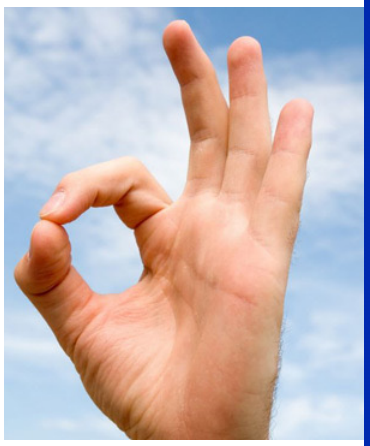
Interesting

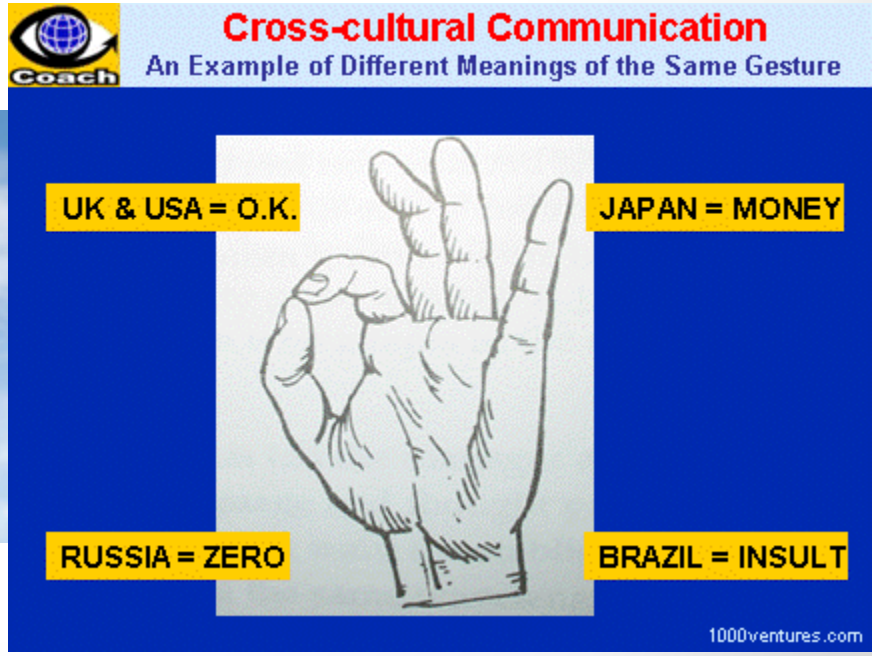
- **Terminology: the difference between a gesture and a manipulation** Posted by Ron in Interaction Design on Sep 6th, 2009 http://blog.rongeorge.com/design/interaction-design/terminology-the-difference-between-a-gesture-and-a-manipulation/



GS - Semaphoric gestures

- "Semaphoric gestures to be any gesturing system that employs a stylized dictionary of static or dynamic hand or arm gestures".



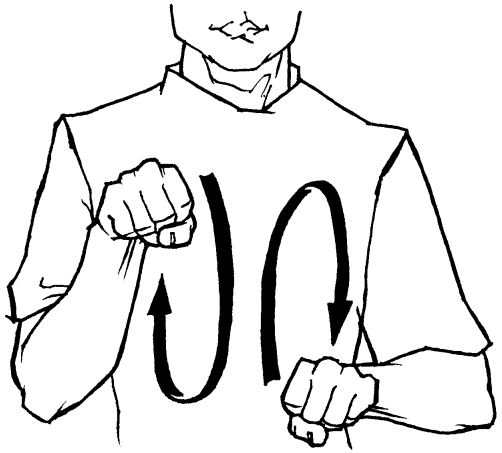


GS - Gesticulation

- One of the most natural forms of gesturing
- Commonly used in combination with conversational speech interfaces
- Gesticulations rely on the computational analysis of hand movements within the context of the user's speech topic
- They are not based on pre-recorded gesture mapping as with semaphores (next).
- Unlike semaphores which are pre-recorded or trained in the system for recognition, or manipulations that track physical movements and positions, gesticulation is combined with speech and does not require the user to perform any poses or to learn any gestures other than those that naturally accompany everyday speech.
- Gesticulations have also been referred to as depictive or iconic gestures that are used to clarify a verbal description of a physical shape or form through the use of gestures that depict those shapes and forms for example

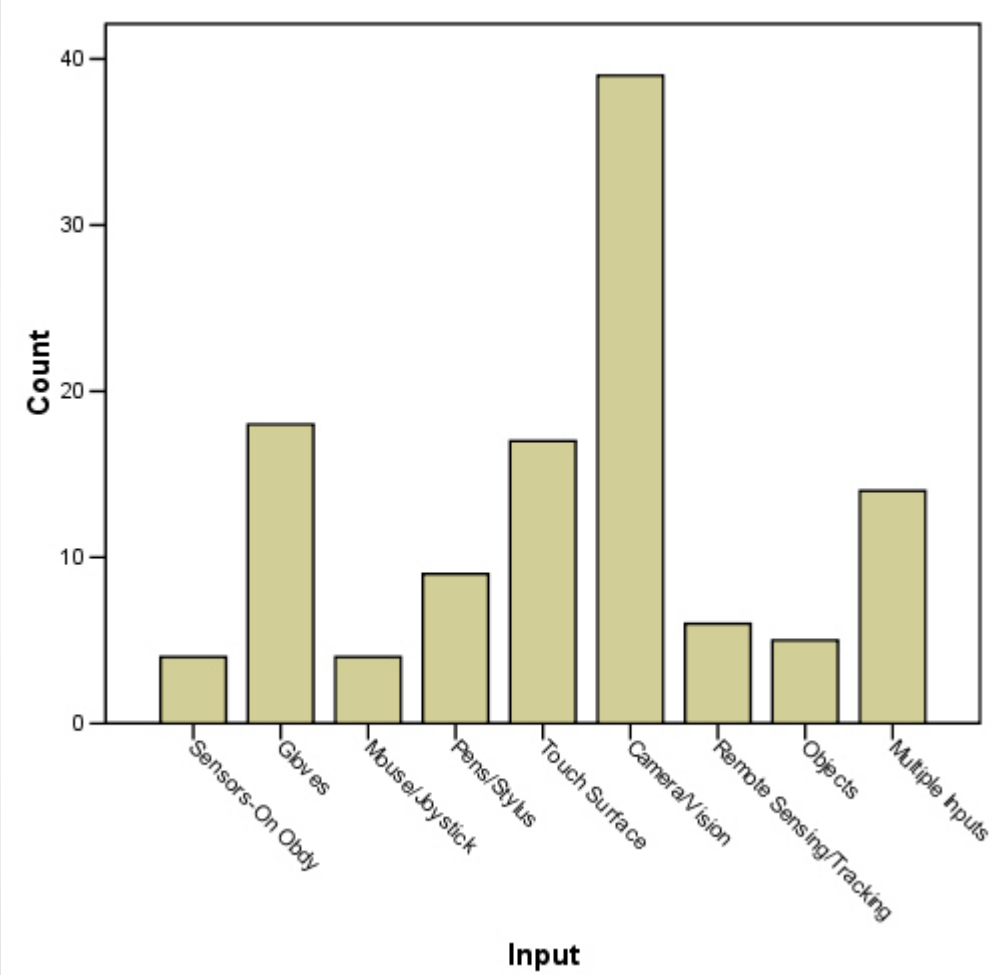
GS - Sign Languages

- Gestures used for sign languages are often considered independent of other gesture styles since they are linguistically based and are performed using a series of individual signs or gestures that combine to form grammatical structures for conversational style interfaces
- In some instances such as finger spelling, sign languages can be considered semaphoric in nature.
- However the gestures in sign languages are based on their linguistic components
- They are communicative in nature, but they differ from gesticulation in that the gestures correspond symbols stored in the recognition system.



Miley Cyrus - **Party in the USA - ASL Song**
http://www.youtube.com/watch?v=QmKnQjBf8wM

Input Technology (IT)



Survey by Karam and Schraefel 2005

perceptual (p) vs. **non-perceptual** (np) input

Maria De Marsico - demarsico@di.uniroma1.it

IT – Non-perceptual input

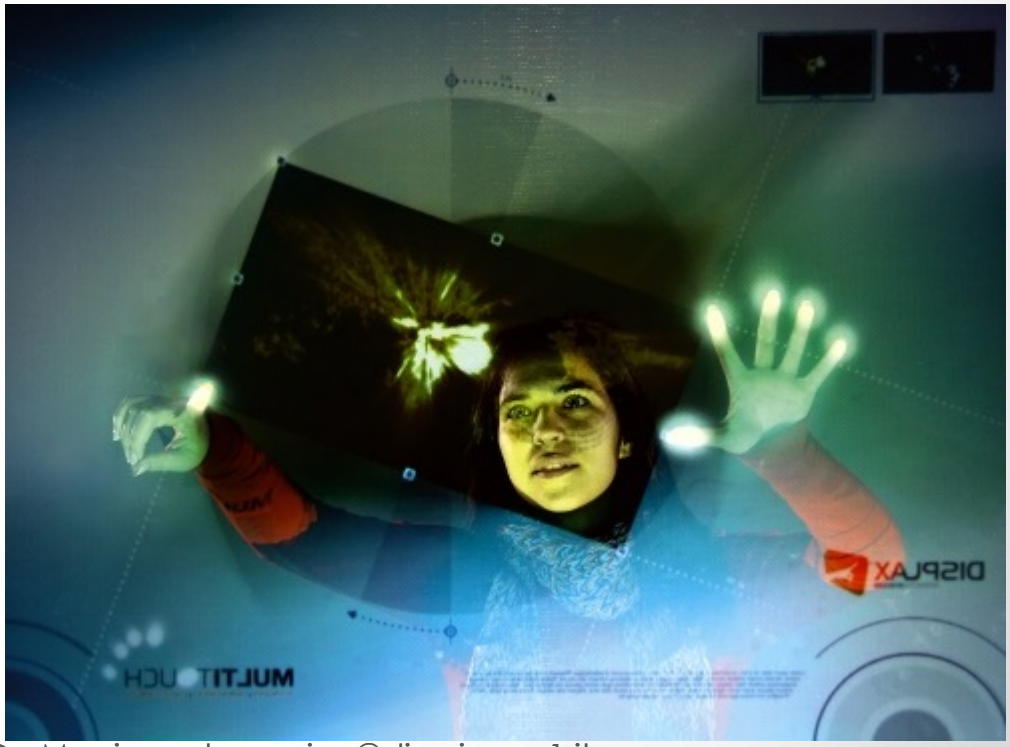
- Non-perceptual input involves the use of devices or objects that are used to input the gesture
- Requires **physical contact** to transmit location, spatial or temporal information to the computer processor.

IT – np - Mouse and Pen Input

- One of the first examples of a gesture based interaction system was Sutherland's SketchPad (1963) which used a light pen, a predecessor to the mouse, to indicate the location of a screen object.
- Gestures using the mouse allow direct manipulation or point and click method of interacting with a computer
- Gestures or strokes of the pen/mouse are translated into direct commands.

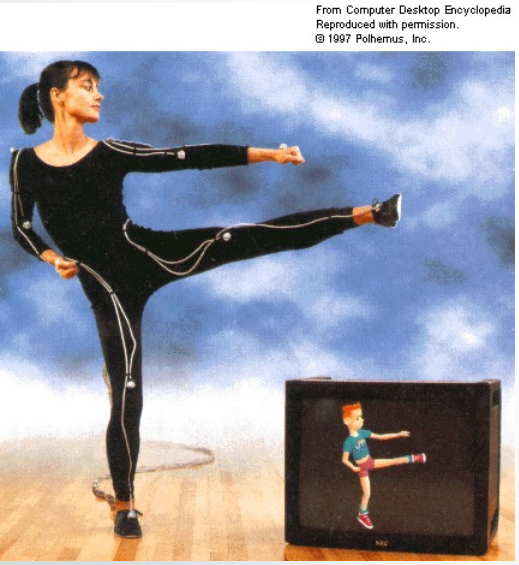
IT – np - Touch and Pressure input

- Touch based input is similar to gesturing with direct input device however one of its key benefits is to enable a more **natural** style of interaction that does not require an intermediate devices like the mouse
- Touch screens
- Mobile computing
- Tablets
- Table top computing



IT- np - Electronic Sensing Wearable or Body Mounted

- These interactions are characterized by the nature of the sensors
- Possible to track space, position and orientation through magneto-electro sensors.
- These sensors (e.g. Polhemus sensors) are one of the primary devices used to directly sense body, arm or finger movements



from http://www.polhemus.com

Maria De Marsico - demarsico@di.uniroma1.it

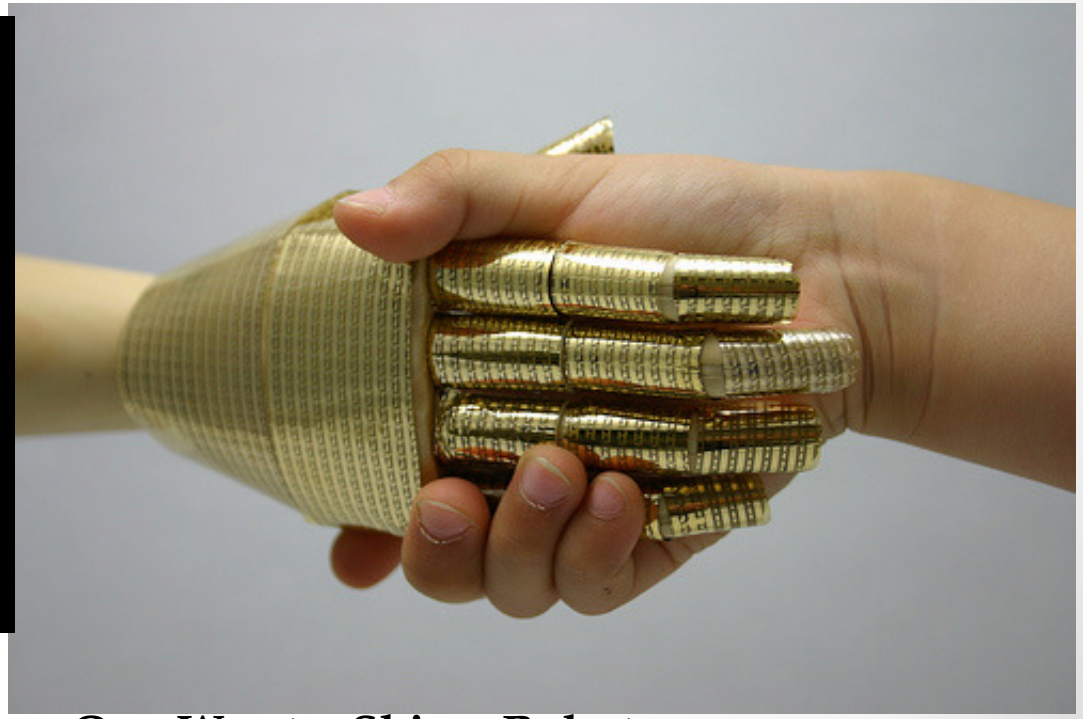
IT- np - Electronic Sensing Gloves

- The use of gloves enables gestures that were more detailed, involving the movement of individual fingers, wrist and hands, to allow a more flexible and accurate gesture recognition
- The glove consists of a tissue glove that is fitted with sensors to measure finger bending, positioning and orientation and may include vibrating mechanism for tactile feedback.

IT- np - Electronic Sensing: Object-

Embedded Sensors and tangible interfaces

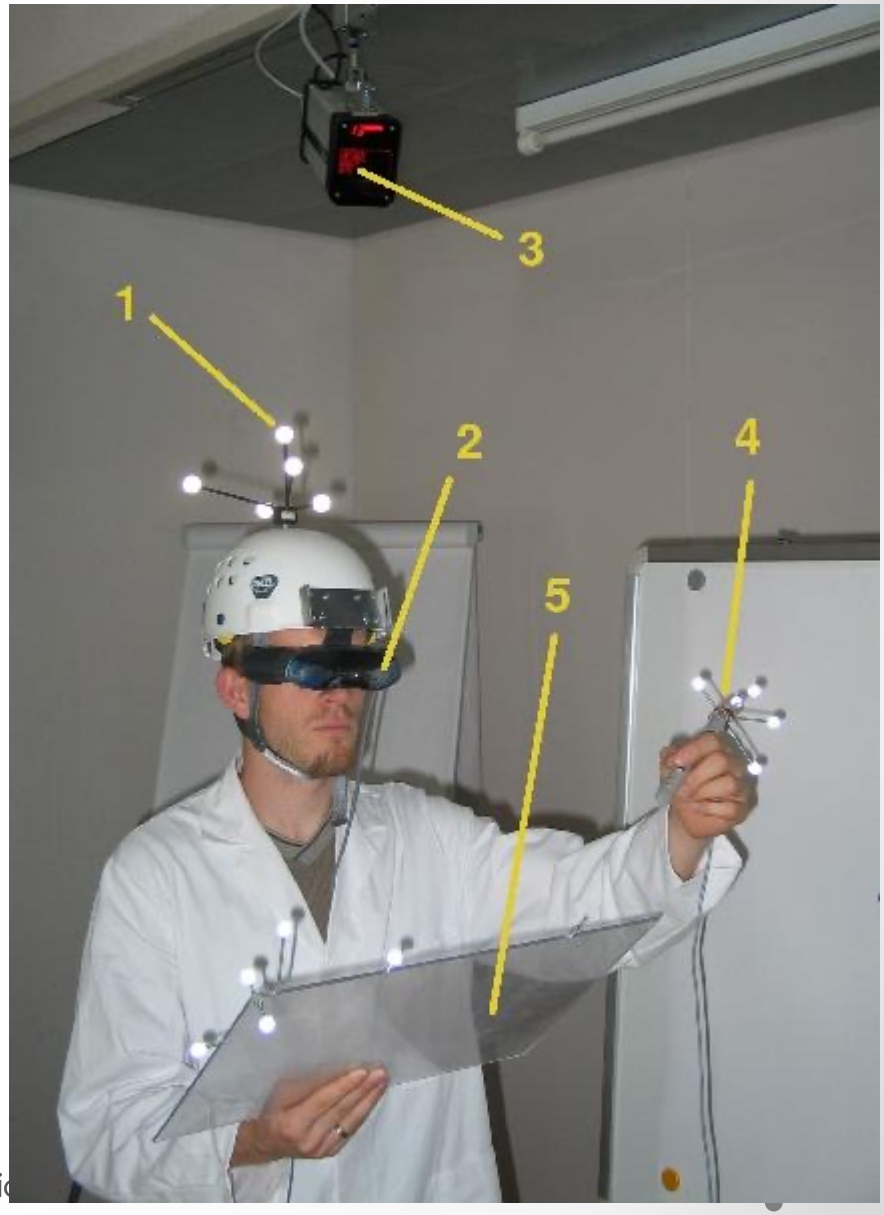
- The sensors move from the user to ambient objects
- Tangible or graspable interfaces



Right image from **There's More Than One Way to Skin a Robot**
http://robots.net/article/2876.html

IT- np - Electronic Sensing: Tracking Devices

- Gesture based interaction can be also performed using infrared tracking devices to detect input.
- The infrared beam is tracked by a camera (**image processing techniques**) and its movements or gestures are translated int0 predetermined system behaviours.

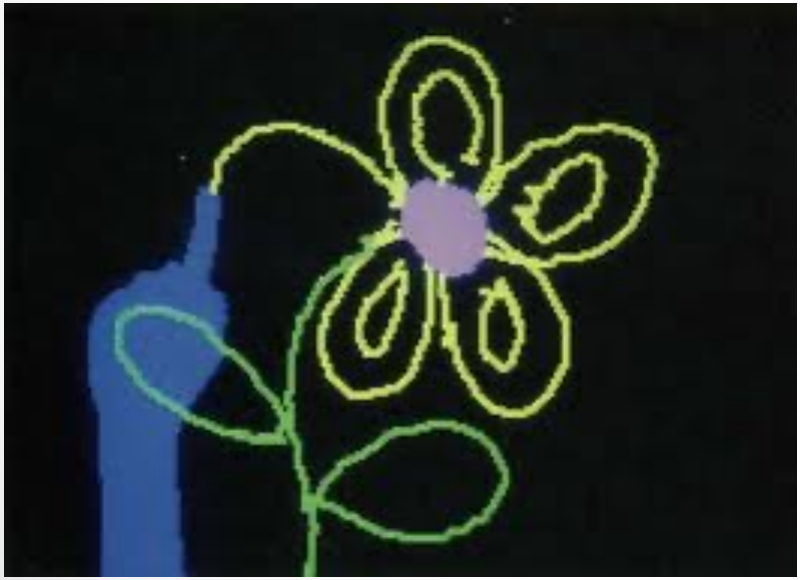


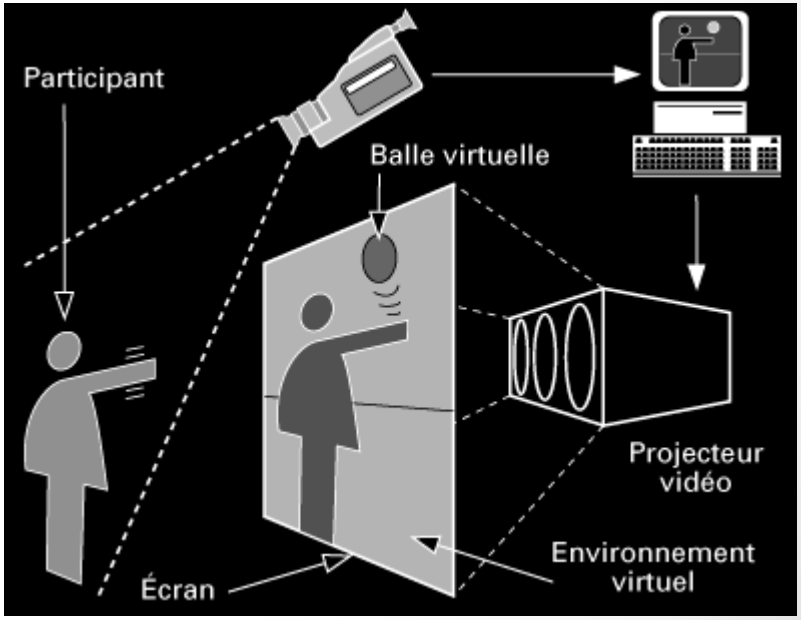
IT – Perceptual input

- Perceptual input enables gestures to be recognized without any physical contact with an input device or with any physical object
- The user can communicate gestures without having to wear, hold or make physical contact with an intermediate device such as a glove or mouse
- Perceptual input technology includes visual, audio or motion sensors that are capable of receiving sensory input data from the user through their actions, speech or physical location within their environment.

IT – p – Computer Vision

- One of the first examples involve using video to recognize hand movements as an interaction mode as seen in Krueger et al's work from 1985 on VideoPlace.
- Krueger's system involved projecting a video image of the user overlaid on a projected wall display. The interaction was based on the user's image coming in contact with or pointing at objects on the display.





Myron Krueger - Videoplace, Responsive Environment, 1972-1990s
http://www.youtube.com/watch?v=dmmxVA5xhuo

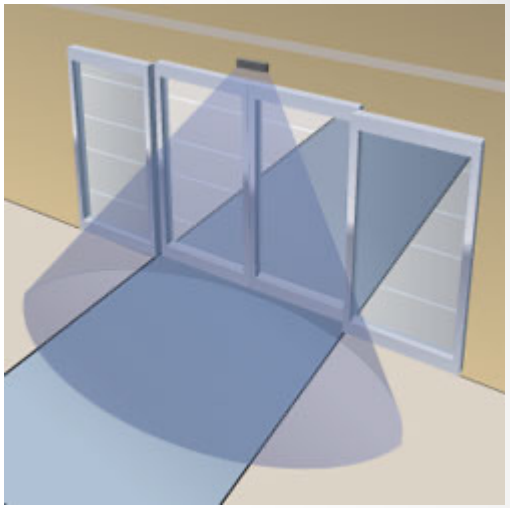
IT – p – Computer Vision

- Computer vision is used for all forms of gesturing **but** poses several research problems in terms of its ability to recognize gestures under changes in lighting and occlusion.
- One work around for some of the sensitivities of computer vision has been to use LED transmitters in combination with cameras. This method allows various gestures to be performed by tracking the interrupted LED output caused by the user performing gestures (few gestures allowed)

IT – p – Motion Sensors

- Passive sensors detect natural radiation that is emitted or reflected by the object or surrounding area being observed. Reflected sunlight is the most common source of radiation measured by passive sensors.
- Examples of passive remote sensors include film photography, infrared, charge-coupled devices, and radiometers.
- Active sensors emit energy in order to scan objects and areas whereupon a sensor then detects and measures the radiation that is reflected or backscattered from the target. RADAR and LiDAR are examples of active remote sensing where the time delay between emission and return is measured, establishing the location, height, speed and direction of an object.
- This form of sensing is used to detect human presence, movement and pressure, enabling full body movements for gesture based interactions.
- Electronic sensors have also been placed on screens for remote sensing of finger movements as an alternative to mouse interactions.

IT – p – Motion Sensors

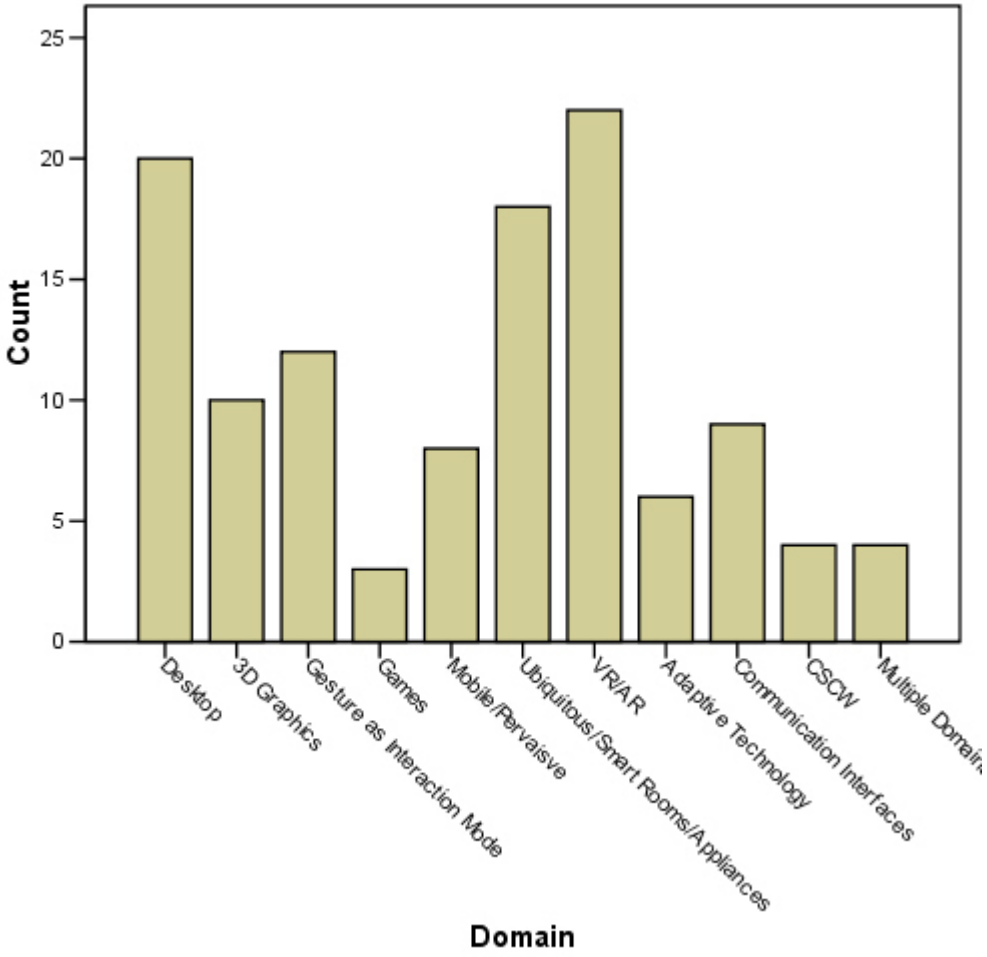


Maria De Marsico - demarsico@di.uniroma1.it

IT – p – Audio

- Tapping location for gesture interaction?
- For large screen displays, identification of the place where the user is going to perform a further interactive sequence
- Perceptual or non-perceptual? (contact is needed)

Application Domain (AD)



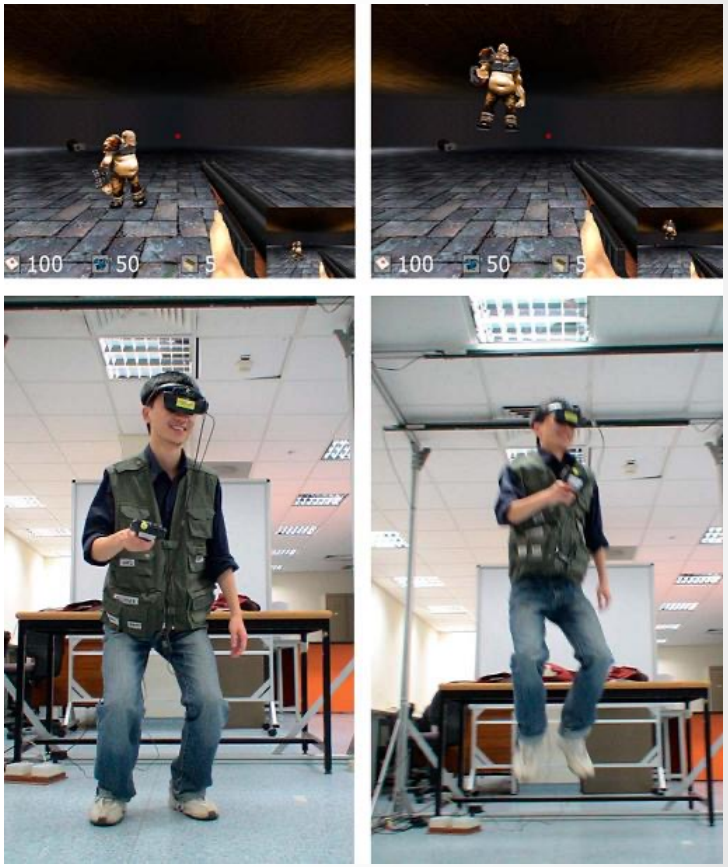
Survey by Karam and Schraefel 2005

Application Domain (AD)

- Virtual and augmented reality (var)
  - Virtual and augmented reality represent one of the largest areas for gesture based interactions. Much of the interactions within virtual reality involve either semi or fully immersed displays although the physical nature of the gestures involved are relatively the same.
- Desktop applications (da)
  - In desktop computing applications, gestures are an alternative to the mouse and keyboard interactions, enabling more natural interactions (e.g. with fingers)
- Three Dimensional Displays
- Ubiquitous Computing and Smart Environments
  - Tangible computing
- Games
- Pervasive and Mobile Interfaces

AD - var - Immersive Interactions and Avatars

- A gesture based interaction style for avatars in virtual worlds consist of **full body movements as a means of modelling and controlling avatar** movements and interactions
- **Sensors** that are fitted on the user's body are used to **track** their body form and movements to create a virtual human on a screen.
- Essentially, **avatars are virtual objects** that are manipulated in terms of their behaviour and their movements within the virtual world.



Augmented reality (AR) and gesture interaction real time 3D characters
http://www.youtube.com/watch?v=1QkBIxGjcMY

AD - var - Manipulation and Navigation in Virtual Worlds

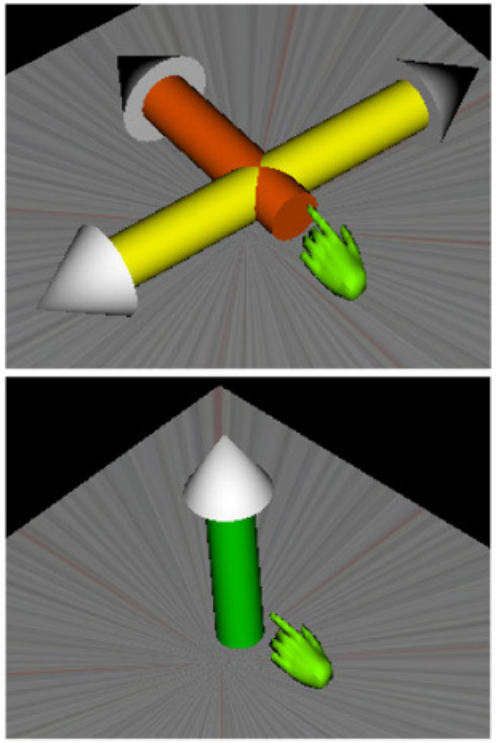


Figure 9: Examples of translation pointers

- When **physical interactions** involve the virtual surroundings and its objects, **manipulative** gestures are typically used
- The detection **of body motion and location** with respect to an area can be implemented using sensors embedded in the environment or worn by the user
- **Movements** in the physical world are **mapped** directly onto the virtual world.
- 3D visualizations that are based on **navigating around objects** within in a virtual world also employ hand gestures as an interaction mode
- Physical manipulation of virtual objects often involve **sensor augmented gloves**.
- During the interaction, the **movements of glove are recreated** within the world as a 3D object allowing a visualization of the interaction within the digital world.

AD - var – Augmented Reality

- In augmented reality, real environment and virtual environment are a **continuum**
- The type of interactions available in augmented reality are very similar to those in virtual reality

AD - var - Robotics and telepresence.

- **Telepresence** and **telerobotic** applications are typically exploited in **space exploration** and **military** based research projects, but also in critical settings.
- The gestures used to interact with and control robots are most commonly seen as **virtual reality applications** as the operator is controlling a robot's actions while viewing the robot's environment through a head mounted display.
- Gestures are used to control the robot's hand and arm movements for reaching and manipulating objects as well as the direction and speed that they are travelling.

AD – da - Graphics and Drawing.

- One of the **first** application domains for which gestures were developed involved graphic style interfaces and was presented as early as 1964
- The pen and tablet style interfaces of the 80's were initially based on tablets or desktop computer screens and typically used a mouse, stylus or puck for the non-touch screens.
- Touch sensors + pressure sensors to determine drawing thickness

AD – da - CSCW.

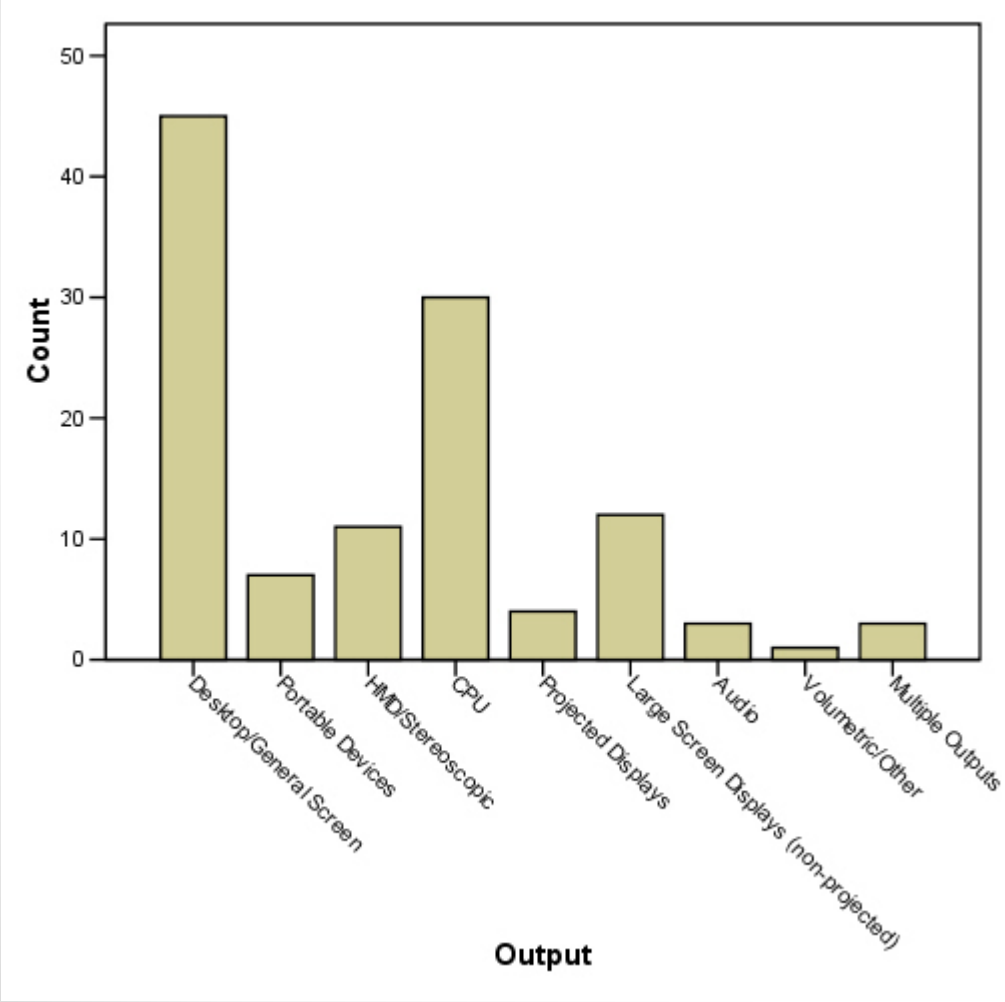
- Table top displays and large screen displays
- Interactions such as sharing notes and annotations among groups

Teliris Brings Touch to Telepresence
http://www.telepresenceoptions.com/2008/06/teliris_brings_touch_to_telepr/

AD- da - Multimodal Interfaces

- Multimodal interactions typically implies the use of gesture in combination with speech interfaces

System response



Survey by Karam and Schraefel 2005

System response

- Visual Output (2D or 3D)
  - Most of the responses that are returned using gestures is based in visual displays.
- Audio Output
  - Audio as output can free up primary attention for more critical tasks such as driving.
- CPU or Command Directed Output
  - The system response of a recognized gesture may not be translated onto any specific output device but rather is used as a control input command to other devices or applications.

Advantages and problems

- Gesture based interactions can provide a more natural form of interacting with computers.
- Gestures enable more natural interactions within virtual reality applications

- The gloves that are often required can be unnatural and cumbersome for the user.
- Computer vision does not allow the same level of accuracy or granularity of gestures that can be achieved with the glove based interaction.
- Markers are a partial solution

Advantages and problems

- **How** a computer can distinguish between individual gestures when they are performed concurrently ?
- This problem involves determining when a gesture **starts**, when it **stops** and when to **interpret** the gestures as a **sequence** rather than individual gestures.

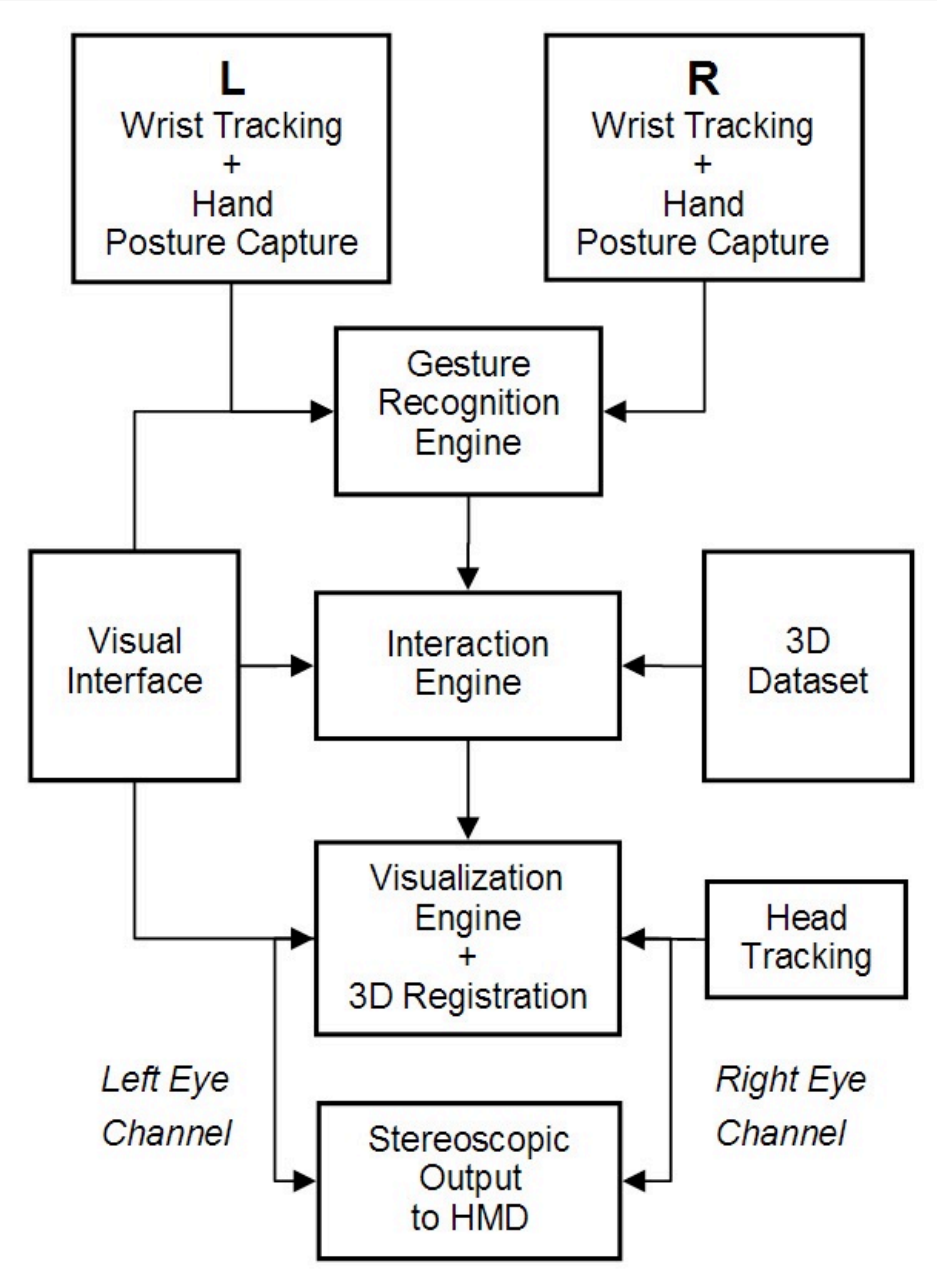
A real example

A Floating Interface Operated by Two-Hand Gestures for Enhanced Manipulation of 3D Objects

+

Virtual keyboard

Developed at University of Salerno


L
Wrist Tracking
+
Hand
Posture Capture
R
Wrist Tracking
+
Hand
Posture Capture
Gesture
Recognition
Engine
Visual
Interface
Interaction
Engine
3D
Dataset
Visualization
Engine
+
3D Registration
Head
Tracking
Left Eye
Channel
Right Eye
Channel
Stereoscopic
Output
to HMD

Modules - Input

- The user **input** module is responsible for user's hands tracking within 3D space and hand posture acquisition.
- An accurate and reliable capture technique based on wireless instrumented gloves and ultrasonic tracking devices I adopted.
- Such choice simplifies the posture/gesture recognition stage, since for example inter-hands and inter-fingers **occlusions** are not an issue : each single finger has individual sensors for flexion and abduction which are unaffected by any other finger.
- As data-gloves do not provide any spatial info, the system relies on a magnetic motion tracking hardware, with six degrees-of-freedom, to detect head and wrists position in 3D space and their rotation on three axes (yaw, pitch and roll).

Modules – Gesture Recognition Engine

- A specific software module checks for particular predefined posture tokens, which trigger associated interaction activities.
- The analysis is performed by a recognition engine based on **timed automata** , able to detect one-hand and two-hands timed posture patterns which are associated to manipulation function.

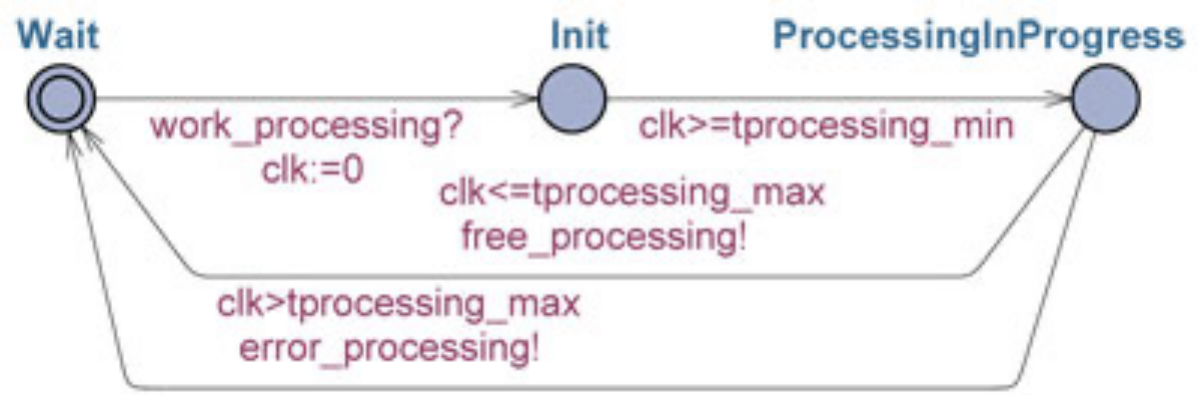
Timed automata

- Timed automata are labeled transition systems used to model the behavior over time of single components in real-time systems.
- Classical state-transition graphs are further annotated with **timing constraints**. Accordingly, a timed automaton performs **time-passage actions**, in addition to ordinary input, output and internal actions.
- In more detail, a timed automaton is a standard finite-state automaton extended with a finite collection of **real-valued clocks**.
- The **transitions** of a timed automaton are labelled with a **guard (a condition on clocks**), an **action**, and a clock reset (a subset of clocks to be reset).

Timed automata

- A **state** of an automaton is a pair composed by a control node and a clock assignment, i.e. the current setting of the clocks.
- **Transitions** are either labelled with an **action** (if it is an instantaneous switch from the current node to another) or a **positive real number** i.e. a **time delay** (if the automaton stays within a node letting time pass).
- Embedding time allows changing the status of involved entities according to **time-based events**.
- This **enhances** the quality of user-system interaction **when a feedback is required in a reasonable time**, or when the time elapsed between elementary actions **can influence the interpretation of their composition**.

Example



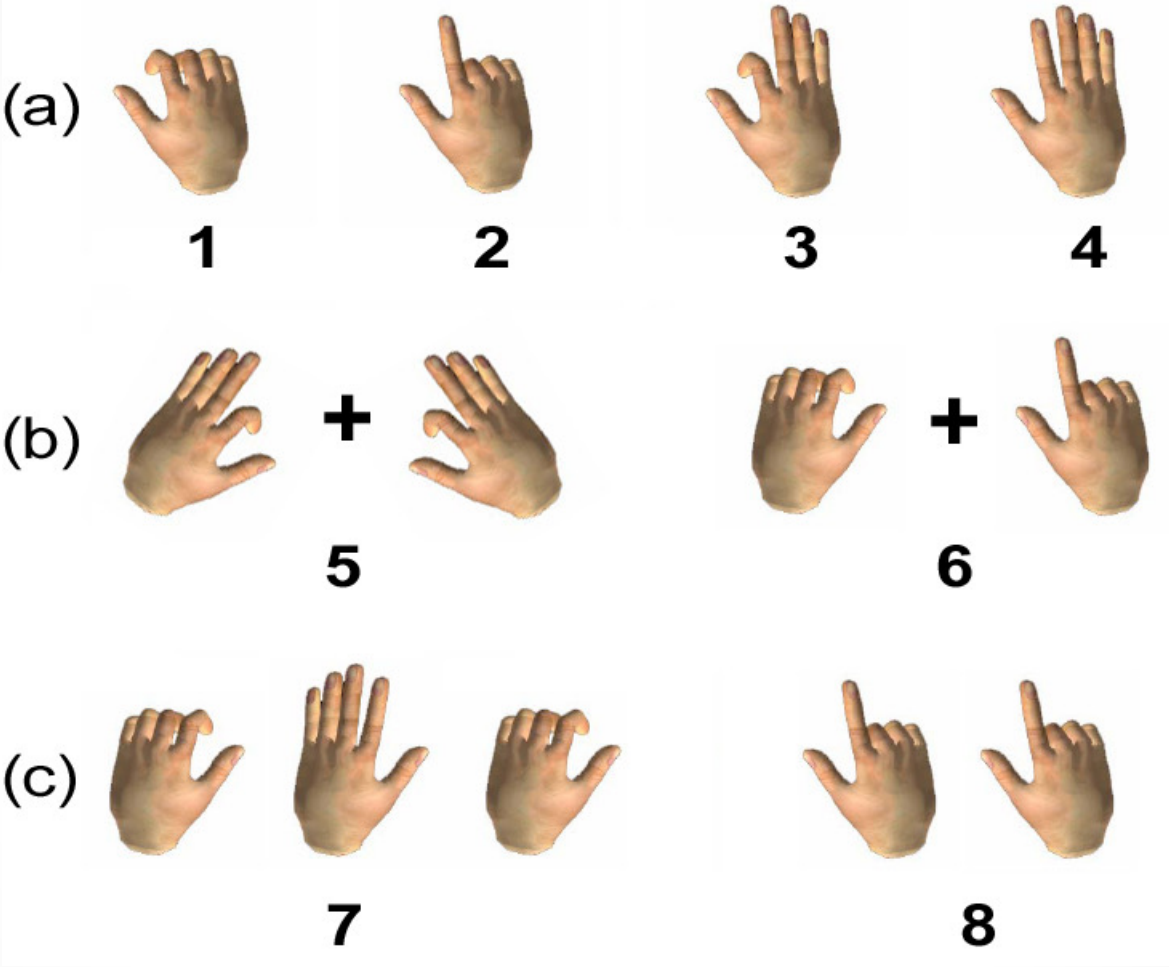
A timed automaton modeling the processing of a task, where *clk* is a clock.

After the reception of a signal *work_processing!*, the automaton spends at least *tprocessing_min* time in the location *Init*.

Then, it sends the signal *free_processing!* if the processing time does not exceed *tprocessing_max*, otherwise, it emits *error_processing!*.

From : Jean-Yves Didier, Bachir Djafri, and Hanna Klaudel. **The MIRELA framework: modeling and analyzing mixed reality applications using timed automata**
https://www.jvrb.org/past-issues/6.2009/1742

Recognized gestures



One-hand (a), two-hands (b) and (c) time-based gestures

Interaction Engine

- Each time a **valid** gesture is fully recognized by the recognition engine, the corresponding vector is inputted to the interaction engine, which exploits a similar architecture based on timed automata and is responsible for any visual interaction allowed by the system, by translating gestures into actions.
- Gestures are evaluated **according to the current interaction status**, so that the same gesture may trigger different actions in different operative contexts (rotation, measurements, landmark assignment, etc).
- Operational modes and manipulation function are selected **via a virtual interface displayed within the field of view** as a frame surrounding the 3D content, and including textual information related to the ongoing operations.

Interaction Engine

- **Visual and acoustical feedbacks** are provided to confirm the "pressure" of a key or the acknowledgment of a particular command, thus reducing wrong operations. If required, interface layout can be hidden at any time via a gesture toggle.
- At present, only a small set of functions has been implemented, allowing to rotate/move the object, to place landmarks over its surface and to take distance measurements between landmarks.
- Object pan is conventionally achieved with any of the two hands. On the other hand, object rotations in 3D space fully highlight the advantage of two-hand gestures.

Visualization Engine

- 3D models of reproduced objects, as well as the virtual interface, are processed by the visualization engine, which is responsible for real time transformation and stereo rendering of 3D scenes.
- These data are processed to transform the virtual content as seen from the user's point of view, and coherently with a 3D model of surrounding environment. Such crucial task is referred to as 3D **registration**.

Maria De Marsico - demarsico@di.uniroma1.it

Hand Gesture Interaction Technology PC Computer Webcam - DealExtreme
http://www.youtube.com/watch?v=pwhSO-Z1Yq8

3D Gesture Interaction
http://www.youtube.com/watch?v=yqstD5GjZEQ&noredirect=1

Gesture processing via MediaPipe

From https://developers.google.com/mediapipe/solutions/vision/hand_landmarker#:~:text=The%20hand%20landmarker%20model%20bundle,by%20the%20palm%20detection%20model.

- The Hand Landmarker uses a model bundle with two packaged models:
  - a palm detection model
  - a hand landmarks detection model.
  - both mopdels are needed for gesture recognition

- The hand landmark model bundle detects the keypoint localization of 21 hand-knuckle coordinates within the detected hand regions.

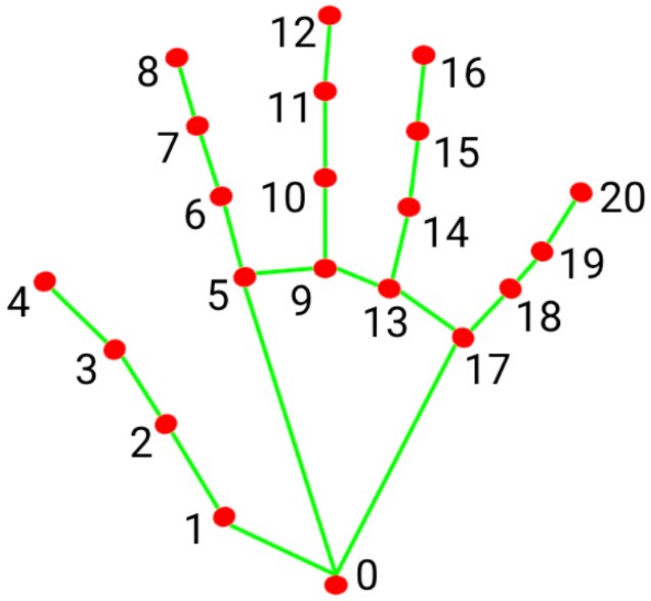
- The model was trained on approximately 30K real-world images, as well as several rendered synthetic hand models imposed over various backgrounds.

MediaPipe Palm & Hand Landmarks

- The hand landmarker model bundle contains a palm detection model and a hand landmarks detection model. The Palm detection model locates hands within the input image, and the hand landmarks detection model identifies specific hand landmarks on the cropped hand image defined by the palm detection model.

- Since running the palm detection model is time consuming, when in video or live stream running mode, Hand Landmarker uses the bounding box defined by the hand landmarks model in one frame to localize the region of hands for subsequent frames. Hand Landmarker only re-triggers the palm detection model if the hand landmarks model no longer identifies the presence of hands or fails to track the hands within the frame. This reduces the number of times Hand Landmarker tiggers the palm detection model.

Hand Landmarks



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Readings

- V.I. Pavlovic, R. Sharma, T. S. Huang (1997). Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review http://www.cs.rutgers.edu/~vladimir/pub/pavlovic97pami.pdf

- Maria Karam and M.C. Schraefel (2005). A taxonomy of Gestures in Human Computer Interaction http://eprints.soton.ac.uk/261149/

- C. E. Swindells (2000). Use that there! Pointing to establish device identity ftp://cieedac.sfu.ca/ftp/fas-info/fas-info/pub/cs/theses/2002/ColinSwindellsMSc.pdf

- Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., and Ansari, R. 2002. Multimodal human discourse: gesture and speech. ACM Trans. Comput.- Hum. Interact. 9, 3, 171–193. http://web.media.mit.edu/~cynthiab/Readings/Quek-p171.pdf