	IST-2002-002114 – ENACTIVE NETWORK OF EXCELLENCE WP4b STAR in Action and vision fusion
Reference	EI_WP4b_DLV1_INPG2_040930
Title	Action Processing
Object	D4b.1 : State of the art in technologies for fusing action and vision Deliverable D4b.1, due to 2004, 30th september
Authors	Luciani Annie, Courroussé Damien
Participants	WP4b participants
Dissemination	WP4b participants
Nb. of Pages	??? pages
Date of the version	2004, September 29th
Confidentiality	ENACTIVE Internal Restricted Use

1 Action processing

©Annie Luciani, Damien Couroussé
September 2004

1.1 Typology of Actions and link with the vision (common with WP4a)

1.1.1 Action. Haptic. Gesture

The term « action » covers different meanings. It states the result of a physical task performed by the human body as well as the way through which this task is performed. Mary M. Smyth & Alan M. Wing [Smyth, Wing 1984] distinguish three levels in performing an action : action refers to what it is done (« drink a glass, pick a pencil »), movements refers to how it is done (the movement, with what movement the glass is drunk) and skill refers to the quality of the movement (how the movement is). A specific action can be acted by several movements. A movement can be colored by several skills. These precise definitions do not correspond easily to the current uses of such terms. Action is often used to name all these features, movement is more general that the movement in the action, and skill refers also to the ability to do something.

Thus, confronted to the multiple and sometimes contradictory meanings of « action » and of « haptic » [Pasquinelli 2004], we will called :

• « gesture » in its usual general meaning, to name all what it can be done by the human physical body, whatever the performed objective, rather than « action » or « movement ».

• « gestural channel » all the sensory-motor apparatus composed of all the physical means, through which the human physical body interacts with the physical external universe: hand, body equipped with all its mechanoreceptors and all its actuators. Gesture is then these human biomechanical sensors-actuators during a physical performance.
• "gestural action" as the motor part of the gestural channel involved in the gestural performance. It involves all the physical components (articulated skeleton and muscles) of body.
• "gestural perception" as the sensory part of the gestural channel. These terms (gestural channel, gestural perception, gestural action) are used to avoid the detailed description of each sub-means (subset of sensors, subset of motor capabilities) as well as the human perceptual and/or cognitive results of the use of these means.

1.1.2 Human-world interaction : three basic functionalities

Having in mind that sensorial events are necessarily produced by physical objects, the sensory-motor relation between humans and environments can be splitted in three basic functions :
• Epistemic function : the function to know the enviroment, from which humans are informed on the environment , by the environment,
• Semiotic function : the function from which humans inform the environment.
• Ergotic Function : the term "ergotic" has been coined by C. Cadoz, [Boissy 1992][Cadoz 1994] to state the relation during which the humans ands the physical worlds are physically interacting, caracterised by the fact that there is a exchange of physical energy between them. The term « Haptic » is often used to state this function. Unfortunately, as stated by E. Pasquinelli [Pasquinelli 2004], this term covers several meanings underlying several different points of view. Conversely, the term « ergotic » coming from « ergos », meaning « physical work, energy », represents clearly the principal property of such function.

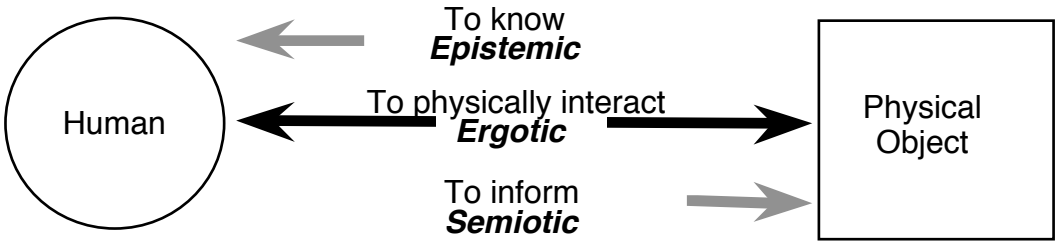


Figure 1. The three functions of the human – environment relationship

Generally, the epistemic function is conveyed by the perceptual apparatus. It is supported as well by the proprio-tactilo-kinesthetic apparatus (mechano and tactile receptors), as the vision and the audition apparatus. Thus, we can speak on the epistemic function of the audition, similarly than the vision as Nivedita Gangopadhyay [Gangopadhyay 2004] proposed with her formal definition of epistemic seeing, and the epistemic function in the haptic sensory modality considered only in its perceptual side, as stated by Hatwell and al. in Touching for knowing [Hatwell, Streri, Gentaz, 2003].

The semiotic function is conveyed necessary by the human channels that are able to emit information to the world. Humans are equipped ONLY BY TWO emitting channels : his/ her mechanical body producing gestural actions (body, arm, hand, face, etc…) and his/her vocal apparatus producing aero-acoustical motions. Some gestural actions aim to transmit pure information (and not energy) to the environment. That is the case of gestures that accompany the speech, the sign language of the deaf-mute, the musical conductor gesture, the gesture of pointing a target with the finger or to point an icon with the mouse, of moving around an object (walking, etc), the cutaneous touch without movements of muscles and joints, of pulling a infinitely light object. Thus the semiotic function is composed of the speech and all the gestures (or motions) the body is able to produce freely (i.e. without any exchange of energy with the external world).

The ergotic function intervenes when physical energy is exchanged during tne interaction between humans and physical world. A specific ability of the gestural channel is to handle directly the matter: to mould it, to transport it, to break it, to cut, to rub, to hit, etc. The hand, (and the all physical body) is in contact with the matter, transmits physical energy to the matter. It applies (and respectively causes) forces, displacements, deformations to the objects and these one react to the human body, resisting to its energetic transfer, and retroacting a part of it. This type of ergotic interaction, aims not only to inform the external world and to be informed by it, but to transformed it. That is possible through a specific property of the gestural channel to be intrinsically bi-lateral: to act on and to perceive, in a inseparable way. During ergotic interaction, and simultaneously of the energetic exchange, humans know (epistemic function) and inform (semiotic function).

The following figure (figure 2) draws all the human – environments interaction channels according to the proposed typology of interaction functions.

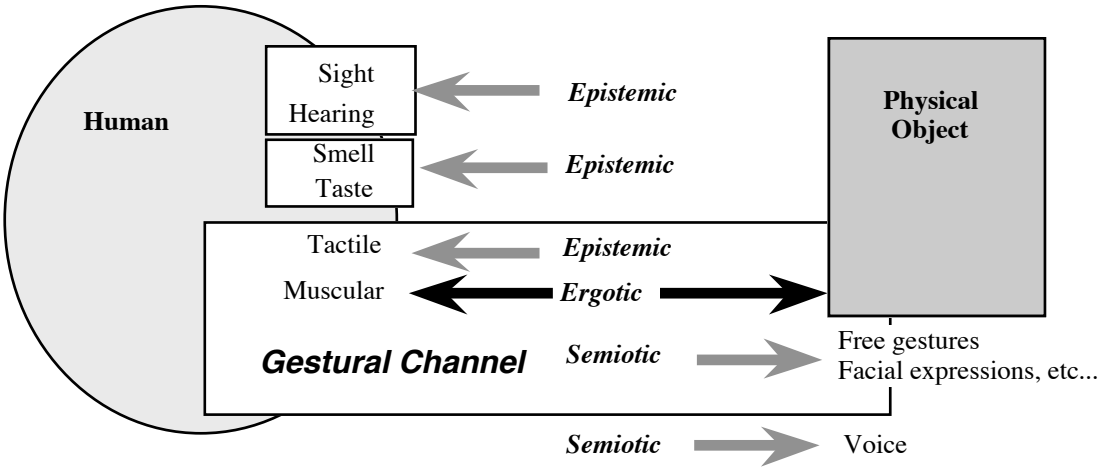


Figure 2. The human interaction channels

In the following, we don't consider the epistemic senses of taste and smell (ambient and localized chemical contact) because of their low relationship with gestures.

1.1.3 The action – perception interaction loops

From the typology of functions in the human-world interaction, several ways of action – perception loops can be declined according to a non redundant criteria of : the existence or not of an energy continuum (or energetic consistency) between the gesture and the perceived phenomena.

1.1.3.1 Pure epistemic-semiotic loops

There are loops linking two grey arrows in figure 3 (sub components of figure 2), in which emission of information from the human subject (to the world) and reception of information by the human subject (from the world) are correlated but without energy exchanges:

- from the semiotic gestural action to the epistemic seeing,
- from the semiotic gestural action to epistemic hearing.
- From the semiotic gestural action (free gestures, facial movements, etc.) and epistemic gestural perceptions, as in cutaneous touch in which there is any noticeable muscular energetic activity in the result of the performance the result of the action. In such actions, the muscular energetic activity can be neglected, or mediated by tools that decrease it without no noticeable loss in the performance of the task.

• From voice to seeing and hearing

Examples are: pointing an object, moving to see or to hear, reading, navigating in a data base or in a virtual environments by means of non retroactive sensors as sticks, mouse, triggering a sounding signals from a non retroactive sensor, selecting an object or an icon, conducting an orchestra, etc. In such action-perception loops, the perceptual result depends on the action but the physical states of the interacting bodies are not modified during and by the interaction process. There are not action-perception loops aiming to act on the world. Mainly there are rather « exploratory activities » oriented to the knowledge of the world or symbolic activities oriented to symbolic constructions.

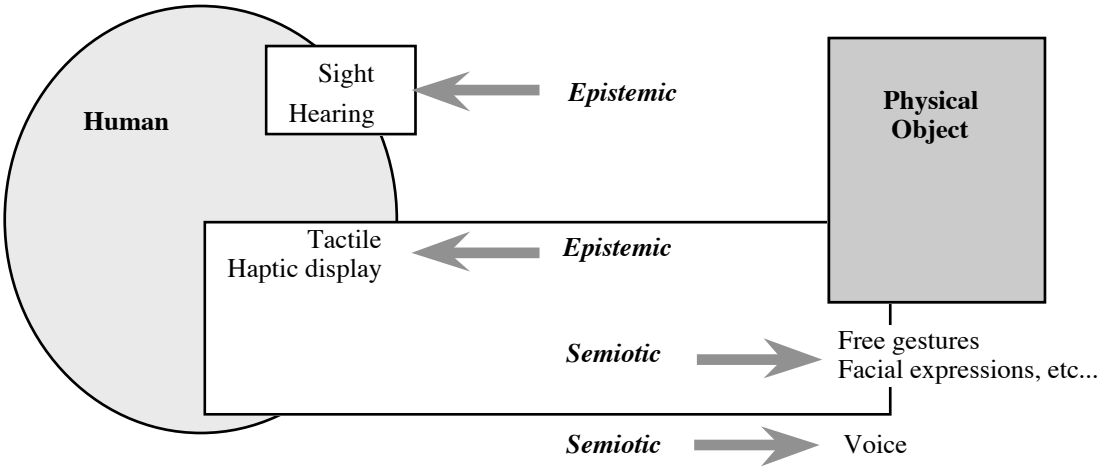


Figure 3. The pure semiotic-epistemic interaction loops

1.1.3.2 Instrumental loop: ergotic interaction and multisensory epistemic feedbacks

It is not sufficient to loop grey arrows as in the pure epistemic-semiotic loops described above to obtain the characteristic property of the ergotic function that is to underlie energy' exchanges (Figure 1, black arrow) between the interacting bodies during the interaction. This means that the ergotic interaction can be clearly-cut distinguished among the others. This points out the operationality of such term to categorize interactions, and consequently to categorize tools and media supporting interaction between a human and his/her external universe. More precisely, the criteria is not the energy spent by an individual during an action, but the energy exchanged between the two interacting bodies, or from the point of view of the human, the energy transferred from (resp. to) human to (and from) object, that is a necessary condition to physically modify the world. As said before, all the handled activities fall in such category: grasping, pushing, pulling, cutting, throwing, carrying, molding, hitting, rubbing, breaking, displacing an infinitely heavy object, writing.

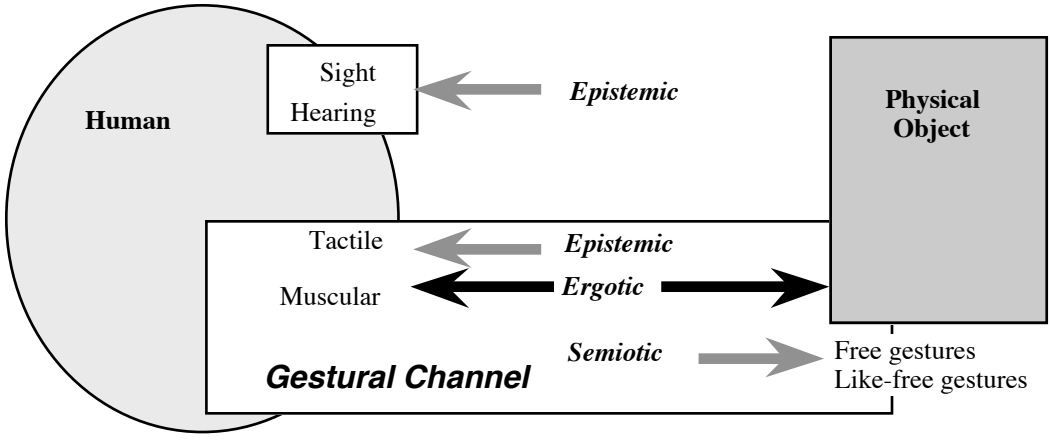


Figure 4. Instrumental Interaction

When one manipulates in such ergotic interaction an object, the physical states of objects (and of humans) are modified by the interaction, exhibiting new mechanical behaviors depending on the interaction (sounds, deformations, fractures, etc…). Thus the sensory epistemic returns (mainly sight and hearing) inform the human of the behavioral answers of the object to the gestural actions. This means that sensory stimuli are not considered by themselves (as conventionally considered by multimodality), but as physical responses that objects do not exhibit without energetic manipulations. These responses encode the coupled system "human body-physical object". They inform the humans of the physical objects and of its physical coupling to the human body. For example, the sound is the encoding of the human / object system during the performance. The visual motion (displacements and deformations) is the encoding of the human / object system during the manipulation, etc. We can state that the physical object transforms the gesture space in auditory (resp. visual) space. The gestural perception informs about the human/object systems that are in contact, etc. The sensorial space is then seen as:

• intrinsically multisensorial : a priori and at least composed of ergotic interaction (with its action and perception part) and acoustical and/or visual returns.

• aiming to know the coupled system object – human.

This means that the object is known (1) through the answers of the matter (that can be sensed) to the gestural actions and (2) all these sensory answers have to be considered a priori as a system encoding the couple human-object and proposing invariants of this system (if they exist) at our cognition and our instrumental ways of control.

We called this typical very common and frequent situation « **instrumental interaction** », with all its declensions (instrumental situation): to dig over the ground, to mould the paste of the bread, to crumple a paper sheet, to play violin, etc.

The instrumental situation is characterized by two necessary features:

- The interaction presents an ergotic component
- The relation between the sensory returns and the gesture exhibits an energetic consistency.

1.1.3.3 Conclusion

The proposed typology is operational in the sense that it proposes one criterium that organizes the different ways of interaction between humans and external world, in non-overlapped categories. This criteria is the existence, or not, of an energy continuum (energetic consistency) between the gesture and the perceived phenomena.

1.1.4 References

[SMY&al, 84] Smyth, Wing. « The psychology of Human movement. » Academic Press, 1984.

[Boissy 1992] Jacques Boissy. Cahier des termes nouveaux. Institut National de la Langue Française, Conseil International de la Langue Française (CILF) and CNRS Editions. 1992. page 52.

[Cadoz 1994] Cadoz C. Le geste, canal de communication instrumental. techniques et sciences informatiques. Vol 13 - n01/1994, pages 31 à 61. 1994.

[Cadoz, Wanderley 2000] Claude Cadoz, Marcello M. Wanderley (2000). Gesture-Music, in Trends in Gestural Control of Music, M. M. Wanderley and M. Battier, eds, ©2000, Ircam – Centre Pompidou, pp. 71-94

[Pasquinelli 2004] Elena Pasquinelli. Some definitions and problems of classification. Enactive WP4b State of the Art

[Gangopadhyay 2004] Nivedita Gangopadhyay. Epistemic seeing. WP4b Enactive State of the Art.

[Hatwell, Streri, Gentaz, 2003] Hatwell Y., Streri A., Gentaz E.. "Touching for knowing : Cognitive psychology of haptic manual perception". John Benjamins Ed.. 2004.

[Luciani 2004a] Luciani A., Urma D., Marlière S., Chevrier J. (2004). PRESENCE : The sense of believability of inaccessible worlds. Computers & Graphics. Elsevier Ed.. Vol 28/4 pp 509-517.

[Luciani 2004b] Luciani A. (2004). Dynamics as a common criterion to enhance the sense of Presence in Virtual environments. Conference Presence 2004. Oct. 2004. Valencia. Spain. To be published.

[Luciani 2004c] Luciani A. (2004). Interaction as exchanged actions and their role in visual and auditory feedbacks. Enactive Virtual Workshop. Enative project.

1.2 Input gesture processing

1.2.1 Motion capture

The Motion Capture systems have been developed to record the movements of human beings. The motion capture technique is essentially used by the animated computer graphics field, either for differed time applications such as the animation of virtual humanoids in movies or video games, or in real time applications, such as artistic performance [Gualtiero, Antonio Camurri, Barbara Mazzarino 2004, WP4b Enactive State of the art] or motion analysis for research purpose.

The motion capture technique corresponds to the recording of the *movements* of an object, not its visual appearance [Menache, 1999]. Therefore, the hypothesis is made that the movement of such an object may be obtained only by the observation of specific points of the recorded object. hence sensors are fixed on these points of the object considered as relevant for this purpose. In the general case of the human body, these relevant points are the joints, such as knee, shoulder, elbow, and extremities of members, such as the top of the head, hands and feet.

1.2.1.1 Main techniques used in Motion Capture

1.2.1.1.1.1.1.1 Optical systems

In that kind of system, which is one of the most used today, the data are provided by (mostly infrared) cameras which record the movements of reflective markers. With the help of two cameras, it is possible then to reconstruct in three dimensions the recorded movements. In most cases, in order to ensure a good reconstruction of the geometrical positions, and to avoid 'phantom' points that come from reflections, it is necessary to use between 7 and 24 optical recording systems. However, the main drawbacks of that kind of system are the fact that the visual field before the cameras has to remain free, and the fact that the amount of post-processing that have to be done after recording may interfere with a real time performance.

1.2.1.1.1.1.1.2 Electromagnetic systems

Magnetic data, if they are not so often used as optical ones are, have some advantages. The main is that orientation of the recorded points may be obtained, and the measure of position is absolute. Furthermore, the magnetic field created by the sensor system defines a zone in which every movement is possible, without the inconvenient of masking the field of one the sensing systems. Therefore, the number of magnetic sensors is often reduced to three. At last, there is no post-processing of data, since the position and orientation of each points are immediately obtained. This allows for real-time applications.

1.2.1.1.1.1.1.3 Mechanical

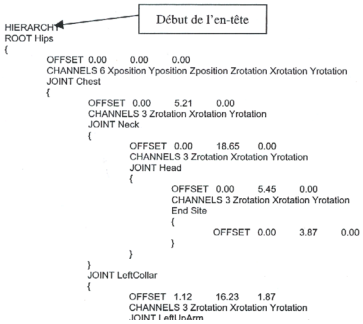
The performer wears a human-shaped set of straight metal pieces (like a very basic skeleton – MetaMotion) that are hooked onto the performer's back. As the performer moves, this kind of exoskeleton is forced to move as well and sensors allocated to each joint provide of measure of the rotation. This technique is interesting when it is not possible to avoid with light or magnetic interference, but it does not allow for an absolute measure of movements, and the displacements of the performer are not immediately obtained. Moreover, the equipment may hinder the performer in its movements.
Other types of mechanical motion capture sets have been developed for specific parts of the body: gloves, mechanical arms, or articulated models such as monkeys or face puppets (PuppetWorks), ("key framing" technique), with which the performer interacts to mimic movements.

1.2.1.2 Post-processing

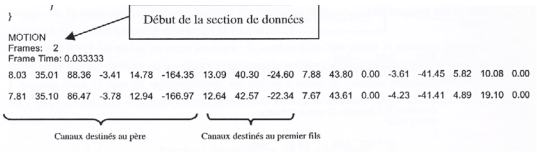
Once the motion data are obtained, it is necessary to export them in a suitable format in order to have them manipulated in computer animation softwares. A *structure* is associated with the motion data, which actually represents the skeleton of the actor, and gives the way the points are related one to each other.
Since the motion capture technique was especially devoted to the recording of human movements, the structure of file formats commonly used permits mainly (and maybe only) the representation of human beings movements.

1.2.1.2.1.1.1 BVA and BVH

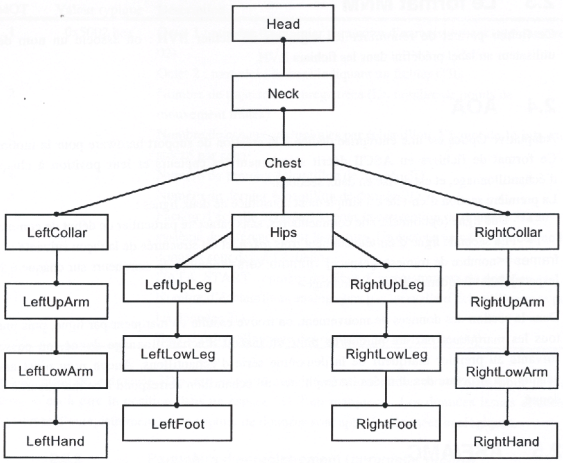
BVH is the acronym for Biovision Hierarchical Data. This file format, which is one of the most commonly used nowadays, has been developed by Biovision. The BVA file format appeared later as a refund of BVH.
The BVH and BVA file formats were especially designed for the movement representation of humanoid forms. BVA files are filled with ASCII information, and contain two sections. The first one is dedicated to the description of the structure of the data. In that section are defined the hierarchy of the skeleton with simple keywords such as OFFSET or CHANNEL. In this section are defined the positions of each *segments* to the origin or to the upper segment is the skeleton hierarchy. The second section is composed of the raw data describing the decomposed coordinated of each segment declared in the first section. The data are described frame after frame, which means that it is possible to obtain the whole configuration of the skeleton at a given time.
One of the main drawbacks of that file format is that it does not allow for an absolute position, nor it allows for explicit rotation information. It is therefore devoted to inverse cinematic processing, but does not allow for a direct interpretation of the data.



The BHV file format - beginning of the first section



The BHV file format - beginning of the first section



An example of hierarchical structure for motion capture cata

1.2.1.2.1.1.1.2 ASK/SDL

This file format is a variant of the BVA file format. The ASK (Alias Skeleton) file contains only the information related to the hierarchy of the skeleton, with absolute coordinates.
The motion information is included in the SDL file. It allows many supplementary information, for example for the description of the scene.

1.2.1.2.1.1.1.3 AOA

Adaptative Optics Associates is dedicated to the creation of hardware support for motion capture. The file format created by AOA is written in ASCII, and allows for very simple manipulation.
The first section is a simple header composed of two lines, which includes comments, the number of frames, the number of markers per sample, and the sampling frequency. The second section contains the motion data. Each line contains one sample of each of the markers. The interest of this format remains in its simplicity: in the second section, the association of a data to a marker is implicit.

1.2.1.2.1.1.1.4 ASF/AMC

This file format was developed by Acclaim Inc., which is involved in video games. It has then been redeveloped by Oxford Metrics (Vicon Motion Capture Systems) when it was put in the public domain.
This format is composed of two files, the first one (ASF file format) for the description of the skeleton, and the second one (AMC –Acclaim Motion Capture– file format) for the raw data information. Its interest remains in the fact that it is possible to associate one skeleton files to many collections of motion capture data describing the same performer in as many motion capture sequences.

1.2.1.2.1.1.1.5 BRD

This file format is dedicated to the unique usage of the Ascension Technologies “Flock of Birds” motion capture system, developed by LambSoft. It only allows the recording of data arising from a magnetic system.

1.2.1.2.1.1.1.6 HTR

HTR is the acronym for Hierarchical Translation-Rotation. This file format was developed as a native format for the skeleton of the Motion Analysis software. It was made as an alternative to the BVH file format in order to make up for its main drawbacks.
The HTR file format contains four sections contained in one single file: header, hierarchy and name of the existing segments, intial position and motion data. In addition to the information available in the BVH file format, the HTR first section allows to define the number of segments, the order of

disposition of the Euler angles, the calibration unities, the rotation unities, the gravity axis, the default axis along which are aligned the skeleton segments, and a globa scale factor.
The interest if this file format, compared to the BVH, is that the hierarchical information is provided independently from the characteristics of each segment, which simplifies its reading and treatment.

1.2.1.2.1.1.7 National Institute of Health 3D

The C3D file format is born from a common need of the research laboratories working on Clinical Gait, Biomechanics and Motion Capture studios to have a universal format for the exchange of motion information. It is thus a public domain file format.
The required properties of this format were the following:

- The ability to store 3D and analog data in an unprocessed form. It is not essential that data is stored without processing, but the format needs the ability to support raw coordinate and analog sample data.
- Preserve information that describes the physical design of the laboratory such as EMG channels used, force plate positions, and marker sets etc.
- Store Trial information relating to the circumstances of the test session such as sample rates, filenames, dates, EMG muscles recorded etc.
- Store Patient information - name, age at trial, with physical parameters such as weight, leg length etc.
- Store calculated analysis results such as gait timing, cycle information and related information.
- Flexibility and compatibility - it must provide the ability to store new information without making older data obsolete.
- A public specification and format description so that anyone can access data without depending on a manufacturer for information.

The C3D file format is therefore a standard that we should classify apart from the other formats, because of its ability to stock a great amount of data and its capacity to comply to very different needs and skeleton structures.

1.2.1.3 Towards standards for motion formats

The MPEG-4 norm, especially developed in the aim to provide an object oriented coding format for the audiovisual scenes, proposed a relatively elaborated model of humanoid objects. This kind of model was optimized in order to allow low transmission costs but a good fidelity for the reconstruction of the scene (i.e. a good replica of human bodies and of their movements) at the reception side.
The MPEG-4 protocol of transmission is pretty inspired by the general form of most of the file formats presented above: the structure and the initial position of the animated bodies is transmitted at first; this first phase provides information about the skeleton composition, surface and texture information for the reconstruction of the whole animated scene, etc. The second phase of the transmission then consists in an update of the pre-defined objects.
The standard MPEG-4 definition of the humanoid structure has the following characteristics:

- 6 degrees of freedom (dof) for the whole skeleton
- A total of 62 dof for the internal movements of the skeleton
- The hand structure, optional, is defined apart, and contains 25 dof
- The skeleton is divided into 29 segments without hands, or 59 segments if we include the hands.

This illustrates the great complexity and the great specificity of the human skeleton structure, regarding to most of the objects included in animated movies, video games, etc.

Therefore, because of the great specificity of human movements and the great difficulty to render them correctly, motion capture techniques are especially intended in the aim of representing human

movements. The motion capture hardware is designed for the recording of human movements, and the software formats reflect this in the way they are conceived.
Furthermore, the relation between the techniques for capturing movements and rendering tools (computer animation software) is based *only on a geometrical purpose*: it means that motion capture techniques deal only with *extensive variables*, i.e. *displacements*, and force variables are less considered in that particular field of computer animation.

We may state that this particular action-vision chain is proceeding of a kind of mapping between the representing humanoid form and the represented object, which is the result of the motion capture process; therefore, the relation between action and vision is opposed to the enactive one since it somehow considers sensed motion as a determined data independent of the behavior of the controlled object, *and therefore do not admit interaction in which such data are influenced by the motion of the manipulated object.*

1.2.1.4 References

[Menache, 1999] A. Menache. Understanding Motion Capture for Computer Animation and Video Games. HandBook. Morgan Kaufmann, 1999.

- Optical systems:

Adaptive Optics Associated: http://www.aoainc.com
Mikromak GmbH: http://www.mikromak.com/
Motion Analysis Corporation: http://www.motionanalysis.com/
Vicon Motion Systems: http://www.vicon.com

- Magnetic systems:

Ascension Technology Corporation: www.ascension-tech.com
Euclid Research: www.euclidres.com
Data glove iReality: www. genreality. com
Polhemus: www.polhemus.com

- Hybrid systems:

MetaMotion: http ://www.metamotion.com
Digits'n Art: http://www.DnAsofl.com/
PuppetWorks: http://www.puppetworks.com/index.htm

- File formats:

Biovision Motion Capture Services: http://www.biovision.com
Format BVIH: http://www.es.wisc.edu/ginphics/Courses/cs-838-l999/Jeff/BVH.html
Format C3D: www.c3d.org
Format IHTR, GTR: http://www.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/{HTR.html, TRC.html}

LIVE Motion Control Platform: http://wscg.zcu.cz/wscg2001/Papers 2001 /R240.pdf
INRTA: *A Global Framework for Motion Capture* http://www.inria.fr/iTrt/rr-4360.html
A proposai for body animation: http://ligwww.epfl.ch/mpeg4/
http://tigwww.epfl.ch/~boulic/mpeg4_snhc_body/body_prop.html

1.2.2 Gesture recognition

We must care the proofreader that this chapter is intended only to give a general approach of gesture recognition, and not to give a detailed view of this complex and still growing field of research, as it does not concern immediately the enactive understanding of gesture. This analytical part might lead to future developments.

In the first 80's, Christopher Schmandt and Eric Hulteen developed at the Architecture Machine Group at MIT a system which allowed the user to control displayed objects on a screen with the complementary help of voice and gesture [Schmandt & al., 1982]. A voice recognition system was combined with a gesture device (ancestor of MoCap devices), which could record the position of the user's finger is space. Sitting on a chair, the user was able to control the display of objects on the screen by pronouncing key-words: "put-that-here"; as each keywords was pronounced, the user had to point its finger at a location on the screen.
With the design of that new kind of device, a user was then able to use speech to control a computer interface, but was also able to interact with a computer interface thanks to a semantic coding associated with gestures.
This opened to new kind of interfaces commonly currently called gesture recognition interfaces.

The primary goal of gesture recognition research is to create a system which can identify specific human gestures and use them to convey information or for device control. The most common applications of gesture recognition are sign language recognition [Ong et al., 2003] [Vogler et al., 2003], device control or interactive approaches of computer interfaces or large VR environments [de la Rivière, 2003]. The general approach of gesture recognition is rooted into the idea of bringing semantic interaction with computer thanks to gesture. Therefore, gesture recognition systems are designed to mainly recognize hand (and arm) and head movements.

The seductive general tendency is to leave the user free of its movements, i.e. contrary to the motion capture, the aim of gesture recognition research is to use human motions as an input of a computer interface without obliging the user to wear a suit or to hold a tracking-tool. Therefore, it is necessary to design systems that have the ability to recognize human presence (static postures) or motion in a video frame.

Although there is no common definition and meaning under the word *gesture*, it is generally admitted in that it encompasses [Brockman, 2003]:

- Static gestures, which are called *postures*.
- Dynamic gestures, which are called to be *motions*.
- Pointing gestures, which are based on a specific location of a limb.

A posture – of the hand for instance – is generally understood as a specific position, or a characterized deformation of the hand and bending of the fingers. In other words, a given specific position of the hand will be said to be a posture if the hand's position can be recognized as one previously defined.

A general overview of a gesture recognition system (Figure 1) might be the following [McGlaun et al., 2003]:

- Video frames are loaded into computer from the recording of a video camera.
- The *segmentation* phase then consists in localizing candidates in the video image for the position of the user's head or hand.
- The gesture features are extracted from the continuous video frame
- Once the motions are recognized, a kind of classification is made in order to determine which gesture did perform the user.

In some applications, the camera is moving as well in order to localize the position of the user's hand [Brockman, 2003].

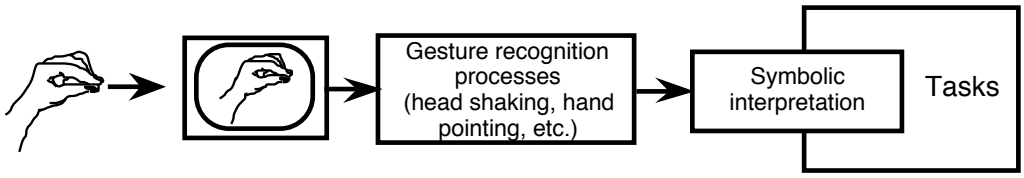


Figure 1. General overview of gesture recognition framework

Many extraction techniques are available for the extraction of gestures in a video frame [Yang, 2002]. *Knowledge-based top-down method*, *bottom-up feature-based method* and *template matching* are based on correlation methods: given a standard face pattern (usually frontal), the correlation values are obtained from an input image and the standard patterns. The existence of a face is determined based on the correlation values. Some other applications, which focus on applying analytical methods for breaking down motion sequences and recognizing patterns, use stochastic algorithms techniques combined with Hidden Markov Models [McGlaun et al, 2003], or Bayesian networks [Ong and al. ,2003]. [Schmidt and al., 2003] based their approach of motion filtering method on a dynamic model with a control system for the arm: the idea is that recognition of human arm motion can be improved by the knowledge of the dynamic control performed by the human motor system. This new approach actually deals with motion recognition technologies mixed with models of human motricity.

The first studies on gesture recognition tried to extract simple components of human movements, such as in sign language recognition or in gesture communication. Such a typical example is [McGlaun et al., 2003], where a corpus of "gesture-words" is constructed by simple head-movements along one degree-of-freedom. The obtained movements are mainly purely rotational movements: moving head on the left, right, up, down, bending left and right, and at last head nodding and shaking. [Brockman, 2003] combines dynamic gestures (i.e. movements) of the hand with "static hand gestures", i.e. hand postures and pointing gestures, i.e. arm or hand predefined locations.

Some recent works succeed now in addressing grammatical inflections of gesture speech [Wong, 2003]. Once the recognition of the "basic blocks" is achieved, it is possible to detect small inflections in gesture, which are more relevant of a subtle meaning. In other words, that is to create relations between the gesture performer and the computer interface that go beyond the simple but now well-established interaction with elementary gestures.

1.2.2.1 References

[1] C. Brockmann and H. Müller. Remote vision-based multi-type gesture interaction. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 198–209. Springer-Verlag Heidelberg, April 2003.

[2] A. Camurri and G. Volpe, editors. *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers*, volume 2915 of *Lecture Notes in Computer Science*. Springer, 2004.

[3] J.-B. de la Rivière and P. Guitton. Hand postures recognition in large–display VR Environments. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 259–268. Springer-Verlag Heidelberg, April 2003.

[4] A. J. Howell, K. Sage, and H. Buxton. Developing task-specific rbf hand gesture recognition. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 269–276. Springer-Verlag Heidelberg, April 2003.

[5] G. McGlaun, F. Althoff, M. Lang, and G. Rigoll. Robust video-based recognition of dynamic head gestures in various domains – comparing a rule-based and a stochastic approach. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 180–197. Springer-Verlag Heidelberg, April 2003.

[6] S. C. Ong and S. Ranganath. Classification of gesture with layered meanings. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 239–246. Springer-Verlag Heidelberg, April 2003.
[7] C. Schmandt and E. A. Hulteen. The intelligent voice-interactive interface. In *Proceedings of the 1982 conference on Human factors in computing systems*, pages 363–366. ACM Press, 1982.
[8] G. S. Schmidt and D. H. House. Model-based motion filtering for improving arm gesture recognition performance. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 210–230. Springer-Verlag Heidelberg, April 2003.
[9] C. Vogler and D. Metaxas. Handshapes and movements: Multiple-channel american sign language recognition. In A. Camurri and G. Volpe, editors, *5th International Gesture Workshop, GW 2003, Genova, Italy*, pages 247–258. Springer-Verlag Heidelberg, April 2003.
[10] R. Watson. A survey of gesture recognition techniques. Technical report, Departement of Computer Science, Trinity College, Dublin, 1993.
[11] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1):34–58, 2002.

1.3 Input-output gesture processing

Input-ouput gesture processing is related to the computer processing of gestures that are bilateral. The case on which gestural action is decoupled to gestural perception (as if we act with one hand and perceive with the other hand- will not be examine here. Thus, the paragraph is composed of two parts :
• Link between the input (gestural action) and the output (gestural returns)
• Link between the gestural input-output loop and the visual output.

1.3.1 Link between the input (gestural action) and the output (gestural returns)

As defined previously, the instrumental interaction contains necessarily ergotic interaction during with there is an exchange of energy between human and manipulated objects. The devices able to convey such coupling between sensors and actuators are necessarily retroactive mechanical transducers as defined in [EI_WP6_DLV1_UPMF_040923]. As said before, such technological situation is new in the context of computerized environments. It was the last to be implemented and its needs for this simultaneously force feedback devices and physically-based models able to produce the forces as answers to the sensed actions.

The process that have to link gestural inputs to gestural outputs is then of a nature deeply different that signal processing (filtering, extraction, reconstruction etc.) used in pure semiotic action as in motion and gesture capture, gesture recognition or expressive gesture presented just before. It is necessarily a computer simulation of a physically-based model, in its large meaning of model that correlates in a energetically consistent way, extensive variables (ex. positions) and intensive variables (ex. forces).

1.3.2 Link between the gestural input-output loop and the visual output.

In such ergotic situation, seing is the way to perceive the object' behaviors that do not exist without such manipulation (deformations, fractures, etc.). The energy communicated by the human to the manipulated object, and retracting on the human body, is the energy that produces such behaviors. What it is seen is the trace, the encoding of the bilateral gestural action-perception loop in a visual effect. We can state that the visual effect is not an effect arbitrarily or metaphorically linked with such actions, but a type of exteroceptive perception of ergotic gestural interaction. Consequently, the relation between action and correlated vision have to be taken in charge (1) by a physically-based model and (2) which have to be energetically link with the model that link the gestural input and output in a consistent way.
Such system is not totally implemented in usual VR platform. [Luciani 1991] [Florens 1998] [Uhl 1995] developed such complete processes in a generic way, that are able to link bilateral ergotic

interaction and epistemic seing of the dynamics of the objects in a generic and physically consistent way.

1.3.3 References

[Florens 1998] FLORENS J.J., CADOZ C., LUCIANI A., " A real-Time Workstation for physical model of Multisensorial Gesturally controlled Instrument". Proceedings of ICMC 1998.

[Uhl 1995] UHL(C), FLORENS JL, LUCIANI(A), CADOZ (C) - «Hardware Architecture of a Real Time Simulator for he Cordis-Anima System :Physical Models, Images, Gestures and Sounds» - Proc. of Computer Graphics International '95 - Leeds (UK), 25-30 June 1995 - , Academic Press. - RA Ernshaw & JA Vince Ed. - pp 421-436

[Luciani 1991] LUCIANI (A), JIMENEZ (S), FLORENS (JL), CADOZ (C) & RAOULT (O), "Computational physics : a modeler simulator for animated physical objects", Proceedings of the European Computer Graphics Conference and Exhibition. Vienna, Austria, Septembre 91, Editeur Elsevier