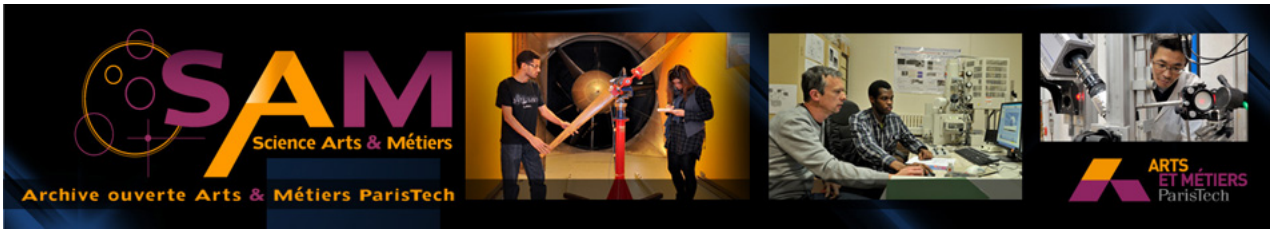
# Is a Time-Of-Flight Camera Better than a Mouse for 3D Direct Selection in Virtual Worlds?

Franck Hernoux, Olivier Christmann, Simon Richir



# Is a Time-Of-Flight Camera Better than a Mouse for 3D Direct Selection in Virtual Worlds?

**Franck Hernoux[1], Olivier Christmann[1], Simon Richir[1]**

**[1] Arts et Metiers ParisTech – LAMPA**
*2, Boulevard du Ronceray*
*BP 93525*
*49035 Angers Cedex 01*
*hernouxfranck@yahoo.fr, olivier.christmann@ensam.eu, simon.richir@ensam.eu*

*ABSTRACT. We present an empirical study on direct selection tasks in virtual environment (VE). The aim is to assess the interest of a new markerless hand tracking system, based on a time-of-flight 3D camera, on a classical mouse, by comparing performances and subjective judgments relatively to utility, usability and immersion. Performances were similar with our system, compared to the mouse, but the perceived usefulness and immersion were judged better. Our system remains lower than the mouse in terms of efficiency and satisfaction. The study reported here, through a simple selection task, demonstrates the interest of this type of camera for real time motion capture. This contribution is a first step and we have to further study more complex task like navigation in virtual environments and object manipulation (moving, scaling, and orientation).*

*RESUME. Cet article présente une étude empirique traitant de tâches de sélection directe en environnement virtuel. Le but est de démontrer l'intérêt d'un nouveau système de tracking de la main sans marqueur, basé sur une caméra temps de vol, vis-à-vis d'une souris. La comparaison porte sur les performances obtenues ainsi que sur les jugements subjectifs des participants, étudiés selon les critères d'utilité, d'utilisabilité et d'immersion. Si les performances sont similaires avec notre système, comparé à une souris, l'utilité perçue et l'immersion ont été jugées supérieures. Par contre, notre système reste inférieur à la souris en termes d'efficience et de satisfaction. L'étude rapportée ici, à travers une simple tâche de sélection, démontre l'intérêt de ce type de caméra pour la capture temps réel de mouvements. Cette contribution est la première étape mais il est nécessaire d'étudier des tâches plus complexes comme la navigation en environnements virtuels et la manipulation d'objets (déplacement, redimensionnement, orientation).*

*KEYWORDS: 3D Camera, direct selection, experimental study, mouse, markerless interaction, hand interaction*

*MOTS-CLES: caméra 3D, sélection directe, étude expérimentale, souris, interaction sans marqueurs, interaction manuelle*

## 1. Introduction

The desktop metaphor, which appeared on Apple computers in 1984, marks the real beginning of the Window Icon Menu Pointer paradigm [Beaudouin-Lafon, 2004], which goes together with the use of the mouse. Since then, many efforts have been focused on improving the quality of graphics software but they base their interaction on the keyboard and the mouse.

For most common uses, (e.g. office automation, web pages, image editing), applications are in two dimensions (i.e. without depth information). The mouse is thus the indispensable device: it is not costly and can have ergonomic design to reduce muscular fatigue. Manipulation of objects in 2D is very simple and common but more and more applications allow «manipulating» content in a 3-dimensional environment: both games and professional applications (e.g. design, modelling, and virtual reality (VR)). These are also increasingly adopted by the general public (e.g. interior configurators) [Rolland et al., 2012].

Interaction in 3D environments is no more with two degrees of freedom (or three with the rotation), but with six: three degrees for the position and three for the orientation. Conventional devices were naturally adapted: combining a movement in one direction with the mouse and pressing a button (keyboard or mouse) allows a different moving. The use of metaphors tends to facilitate understanding and interaction, like manipulating a rolling sphere for the actions of rotation [Chen et al., 1988]. Although it is not certain whether these uses are correct in terms of ergonomics or efficiency (e.g., [Berard et al., 2009]), they are nevertheless widespread. In fact, no interaction device for the manipulation of 3D objects has been widely adopted, except in some very specific domains such as video games. 3D mice are not expensive (around $ 50), but they are not very common. Their field of use is too limited because they only replace partially the mouse and are dedicated to bimanual interaction [TAG, 2008]. It's now possible to provide devices that are suited to tasks in a 3D environment and acceptable to the vast majority of users. Especially since professional users and the general public are more likely to accept news things than a few years ago.

Currently, most existing devices follow the Norman's model (Figure 1 left): it suggests that the user sends stimuli to any device (mouse, keyboard, etc.) that converts the information and sends it to the computer; this one returns visual data to the user. This model was also enriched (gray arrows) by [Nedel et al., 2003] by adding feedbacks from the computer to the device, and between this last and the user when using haptic devices.

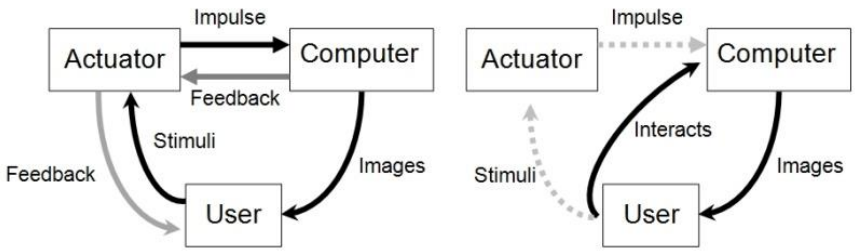



**Figure 1.** *Norman's model, enriched by [Nedel et al., 20003] (left) and our model (right)*

It is now possible to modify a bit this model (Figure 1 right) in the context of VR. Actually, a technology aimed to make the interface "disappear" to allow "natural" interaction for users is often presented as the final outcome of direct manipulation interfaces. This is called behavioural interface [Fuchs and Moreau, 2003]. The sense of immersion and presence can be enhanced by the fact of having no equipment to wear and make the system transparent to the user [Winkler et al., 2007].

How could we interact in VE with the same ease than in reality, without using invasive or intrusive interfaces? The emergence of 3D cameras provides a first answer to this question. Although our work is prior, the Microsoft Kinect is a strong signal in this direction. However, the main issue is to develop new interface and application. The purpose of our work is to report the design and the evaluation of a new solution to capture hand movements without sensors, and thus make possible a real-time 3D transparent interaction for the user. The aim is to provide an effective, comfortable, accurate and efficient system (in terms of accuracy while minimizing the

time of use) which provides a real added value to the user. The system is understood here as a combination of technology, computer vision algorithms and interaction modalities.

The question of the superiority of such a system seems obvious especially compared to a mouse. In fact, there is no proof of the value of a time-of-flight camera for direct selection regarding qualitative or quantitative criteria, in spite of promising properties. This article aims to demonstrate the value of such devices through an experiment comparing a mouse with a 3D camera for a simple task in a VE (selection of objects). The second goal is to show that classical 2D mouse is not anymore sufficient for manipulating object in 3D. First, we present a short overview of the hand, to demonstrate its richness and its complexity. Afterward, we present an overview of current means of interaction, with a particular focus on computer vision based gesture recognition. Then we describe the experiment we conducted with 71 participants and discuss the results. We conclude with the prospects of our work.

## 2. State of the art

### *2.1. The manual interaction, a richness interaction*

The kinematics of the human body consists of more than 200 degrees of freedom (DOF) [Mizuuchi, 2006]. It is thus impossible to consider all these DOF in a real time context. However, a reduction in the number of DOF is not neutral: the construction of a simplified model leads to a loss of realism. It is then necessary to find a compromise between “realism”, simplicity of use and effectiveness in relation to the objective [Lempereur, 2008].

By itself, the hand is able to provide 70% of the human motor skills and thus offers very rich functionalities. We therefore focused on the hands and forearms because they are sufficient in the context of interaction tasks and manipulation of 3D objects. Krout has identified more than 5000 different hand gestures [Krout, 1935]. Through an elaborate network of muscles and tendons the hand is able to perform a multiplicity of tasks.

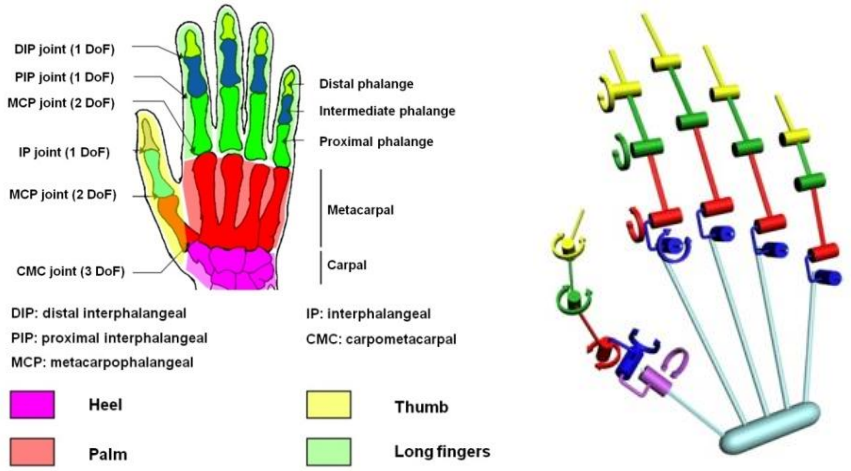



**Figure 2.** *The joints and parts of the hand (left) and mechanical modeling of the hand (right)*

The hand is composed of two parts: the hand itself (palm and heel) and the fingers which have 15 joints (Figure 2 left) allowing 21 DOF (22 if we considers that the carpometacarpal of the thumb has three degrees of freedom instead of 2 [Buchholz & Armstrong, 1992]). Because

of the complexity of its movements, the agreement has not yet been made on a common terminology for defining and measuring its degrees of freedom. The whole hand allows 28 degrees of freedom: 22 for the fingers (if we consider 3 DOF for the metacarpal) and 6 for the wrist (3 rotations and 3 translations). Figure 2 (right) shows the modeling of the index.

Even if a gesture seems natural to us, the complexity of the hand makes it difficult to capture and to interpret for real time interaction in 3D VE. The challenge is to transpose the manual interaction from the real world to the virtual one, with complete transparency for the user.

### *2.2. Motion capture techniques*

The Motion capture describes the activity of analyzing and expressing the human motion in mathematical terms [Bray, 2006]. According to [Menache, 1999]: “Motion Capture is the process of recording a live motion event and translating it into usable mathematical terms by tracking a number of key points in space over time and combining them to obtain a single 3D representation of the performance”.

The capture of hand movements may be based on “software” or “hardware” techniques. For the former, there is obviously a physical interface but a simple one (a camera most of time), the emphasis is mainly put on image processing and more generally on computer vision. The second ones rely on tracking systems that can be electromagnetic, mechanical, optical... We will present a short review of “hardware techniques” and then focus our state-of-the-art on “software” techniques, since time-of-flight camera belong to this category.

#### 2.2.1. Hardware techniques

Hardware systems are the most commonly used. Some systems allow more than six DOF, they are called Very High Degree of Freedom Devices [Hayward, 1996]; among this category, we find exoskeleton, magnetic and optical systems, data gloves and some other few widespread technologies. Exoskeletons (e.g. gloves with Hall Effect sensors) are composed of armatures and joints which are linked to switches, buttons or other complex mechanical systems like pneumatic cylinders or cables held by motors. Magnetic and optical sensors are based on a set of sensors positioned on the user's hand. For the first ones, sensors measure their spatial relationship (position and orientation) relative to a central transmitter which emits an electromagnetic field in a real point. For the second ones, sensors are in fact reflective sensors. 3D positions of each marker are obtained by triangulating shots of at least two synchronized cameras. New algorithms allow to determine the missing positions (due to occlusion) by computing [Herda et al., 2001]. Data gloves (also known as electronic, digital or sensory gloves) are the most common interfaces for detecting partial or total relative movements of the fingers against the wrist. They use optical fiber [Sturman & Zeltzer, 1994], lamellae (or conductive ink) [Kessler et al., 1995] or pneumatic chambers [Sun et al., 2009].

Other systems allow between three and six DOF. This category is balanced between haptic arms, 3D mice, joystick and some hybrid devices. Haptic arms allow a tracking from 3 to 6 DOF, but also apply force feedbacks. 3D mice and the most recent joysticks permit to have at least 3 DOF but their main problem is that the user cannot work directly at 1:1 scale. Finally, some hybrids devices combine several of the techniques mentioned above and can follow the movements of the user's hand according to 6 DOF: the Nintendo Wii controller is the best known and most common example: it combines several sensors (infrared camera, accelerometers, etc.) and a fixed element (sensor bar) which includes infrared LED, allowing a triangulation computation.

Even if they allow a high accuracy combined with high reliability of data, they are expensive and require a calibration phase. All the previous devices allow 6 or more DOF (assuming the use of a sufficient number of sensors), and can be classified as “non transparent” as it is necessary to wear some material (sensors, markers, exoskeletons) or to use a hand-held object (joystick, 3D mouse). None of these devices allows to track the movement in a completely transparent way for the user. For the design of our system, such devices do not meet the constraint of transparency but cameras do. Image processing and analysis is thus necessary to retrieve only relevant information, so we focused our study on software systems.

2.2.2. Software techniques

Lots of research work are focused on recognition and tracking of hand movement without sensors, and rely on one or more cameras. These solutions are based on image processing from one or more video streams. We will only present real time techniques, which allow interaction without latency between user’s actions and visual rendering. These techniques are divided into monoscopic and stereoscopic vision. Figure 3 summarizes the main techniques.

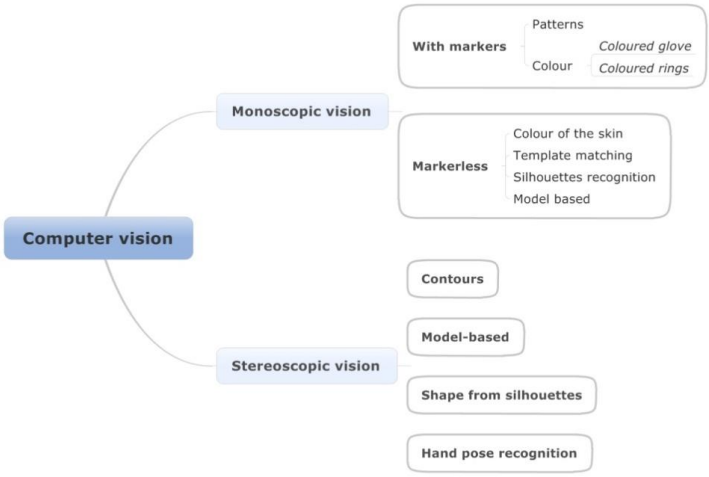



**Figure 3:** *Image processing techniques covered in our work*

2.2.2.1. Monoscopic vision

Monoscopic vision is not sufficient to ensure an accurate 3D hand motion tracking; it is therefore often associated with other techniques. For example, 3D scene reconstruction is possible from visual cues present in the image. However, this requires a learning step and many reference images before working well. Alone, it can’t provide an accurate depth map but can supply additional information when coupled to another system [Saxena et al., 2007]. Among the monoscopic (and also stereoscopic) systems, two main approaches are possible, with or without markers.

Systems based on patterns (binary images) such as the system developed by [Pamplona et al., 2008] to detect the finger’s movements requires cubes with markers (on each side) on 4 fingers and a camera placed on the palm of the user. Coloured gloves [Geebelen et al., 2010] and coloured rings solutions are based on the colour as basic information to track the movements of the hands or fingers. One of the latest methods for coloured glove is used at MIT [Wang, 2009], using 20 coloured areas whose judicious location avoids any ambiguity. This system, which has a framerate of 10-15 fps, remains too limited for real time and requires wearing a glove, but it does not contain technical elements contrary to the hardware systems.

Systems based on the recognition of skin colour [Rautaray et al., 2011] allow to obtain good results but are very sensitive to the brightness which must be constant [Hassanpour et al., 2008]. The “template matching” technique is used in many studies to detect and track the movements of the hand. It may be based on the contours of the hand [Mohr et al., 2009] which are very characteristic for articulated objects. However, a good quality of the contours can be difficult to obtain because of the lighting, camera settings, background colour, shadows, etc. Template matching can also be based on the silhouettes of the hand or colour [Stenger et al., 2006], but this method requires many templates for a single match, which has a significant impact on the computation time and thus on the real-time aspect. The recognition of silhouettes is similar to the template matching: it uses the silhouette user’s hand to position and best match a 3D model. The research project “VTS 3rd Generation” [Tosas, 2006] can track the movements of the fingers using an algorithm that detects the color of the skin and the contours of the hand. Finally, systems based on 3D models are designed to readjust the most similar posture of a 3D articulated model on the image captured by the camera [Ouhaddi & Horain, 1998]. Thus, different models exist such as skeletal models (with line segments) [Dorner, 1994], solid models (with cylinders) [Regh & Kanade, 1994] or deformable surface models (with β-splines or mesh) [Kuch & Huang, 1995].This is one of the most widely used techniques to estimate the postures of the hand from a single video stream. All these works can therefore readjust a 3D model from a 2D image, but unfortunately do not allow a numerical model to be positioned in space according to the three dimensions.

All the techniques based on monoscopic vision can’t give accurate 3D information and tracking of the movements in three dimensions. As for 2D applications, these systems have well proven but they can’t be used to allow 3D interaction in VE. They have to be coupled with other technologies (e.g. sensors positions) or adapted to stereoscopic systems.

2.2.2.2. Stereoscopic vision

The stereoscopic vision is used for the depth reconstruction and is based on the characteristics of human vision. The principle is to use two cameras slightly spaced out (a few centimeters): the differences between the 2 captured images, due to the different viewing angles of the same object, allows to calculate the difference in position of each pixel of the images and to reconstruct the depth map of the scene [Scharstein and Szeliski, 2002]. Many algorithms exist for this reconstruction. Although they are more and more efficient [Di Stefano et al., 2004], they still need to be improved for real time.

A binocular system is sometimes used to reduce occlusions and the number of possible solutions such as in the case of systems based on silhouettes [Kato et al., 2006] or contour [Romero et al., 2008]. [Gumpp et al., 2006] also uses stereoscopy to improve results regarding the placement of a 3D model from the contours of the hand. Systems based on 3D models are also suitable, the model is not matched on a 2D image, but from the points cloud coming from the depth map which is obtained through stereoscopy [Delamarre et al., 1999].

Multi-view reconstructions require several cameras all around the hand. The reconstruction of the 3D volume can be based on different methods. The “shape from silhouettes” method subtracts the background image on each point of view: the volume is then reconstructed from all the silhouettes. This method allows a reconstruction in real time (with a computer cluster most of time) but assumes that the cameras can capture the entire desired volume to avoid any loss of data. 3D surface reconstruction is effective even if the object is not seen by a camera [Michoud et al., 2006]: only a visual hull (the maximal shape) of the object is taken into account. The GrImage platform of the GRAVIR laboratory [INRIA-Rhône-Alpes, 2007], uses this technique and allows 3D reconstructions of persons or objects. It is also possible to insert them into VE, apply real textures and shadows, physics and interactions in real time. Finally, there is the “pose estimation” technique used by [Schlattmann and Klein, 2009] in their 3D data

manipulation system based on three cameras and which not requires any sensor (Figure 4). After the 3D reconstruction, the system detects the pose of the hands to perform the interaction with the 3D object. This system is inexpensive, allows real time and bi-manual interaction, but the user must keep a particular pose of the hand for a few seconds until the system detects the expected action. Another drawback is that the user has to remember the different poses (i.e. the language) which are neither easy nor intuitive.

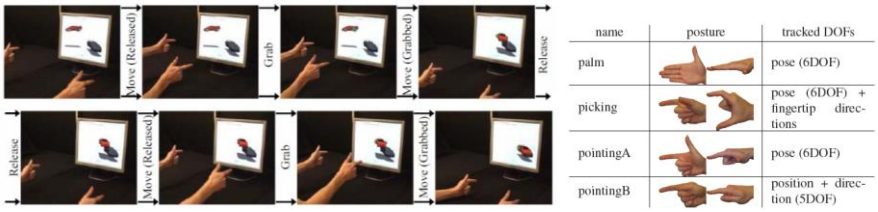



**Figure 4.** *Manipulation of a 3D object based on hand posture estimation*

2.2.2.3. Conclusion

Algorithms used in monoscopic vision do not allow real-time interaction according to the three dimensions. Stereoscopy provides a solution but even if algorithms are more efficient [Stefano et al., 2004], they have to be improved to allow real-time processing with sufficient resolutions. In addition, if the reconstruction of the depth can be performed in real time, it still requires a lot of computation time. Additional treatments necessary to capture and track the movements make, if they are complicated, the final system too slow to be considered real time. It is then necessary to use a cluster of PCs. If lighting problems or background changes remain, stereoscopy induces other complex issues such as parallax and lens distortion for example.

### *2.3. 3D Cameras*

The choice of using one technology over another meets a set of criteria and constraints related to the tasks and application to implement. These constraints are numerous: real-time capture, minimum response time, nature of the movements to track, transparency of the system, size of the workspace, accuracy, resolution, sensitivity to environment or the price. Software solutions can override most of these constraints. The absence (or very low presence) of equipment makes the system seamless (or nearly) to the user that brings interaction in VE close to real environments. In addition, the price is affordable even to individuals, because they require the use of one or two cameras in most cases. However, software solutions are also subject to various disadvantages that may appear redhibitory. Algorithms for image processing, if they are not optimized or too complex, can slow down the system which become unsuitable for real time applications. An interesting solution would be to use a system that would calculate the depth without deporting the computation on the computer. Such systems exist and are known as 3D cameras. An evaluation of the use of these cameras in the field of computer graphics was done by [Kolb et al., 2009]. They are starting to become widespread with the arrival of the Microsoft Kinect. Time-of-flight cameras create depth maps from a single lens (as opposed to the principle of stereoscopy). The functioning is similar to the one used for Lidar scanners with the advantage of capturing an entire scene at once.

These systems can cover distances from 40 cm to over 60 m for some models. The main interest for the capture of the hand is the geo-localization based on the position of the 3D camera, so there is no need for additional trackers, contrary to gloves for example. This type of camera has other advantages such as its relative insensitivity to magnetic or light disturbances (the system is based on infrared light and has the ability to work both day and night, but remains very sensitive to reflective surfaces). Thanks to the depth map, 3D cameras

can isolate the person facing the screen and focus for example on her hands and remove people and objects of the background.

The main drawback of these cameras is their low resolution: 176 x 144 pixels for the SwissRanger SR4000 (Mesa Imaging), 200 x 200 pixels for a PMD Vision CamCube 3.0 (PMD Technologies) and 640 x 480 pixels for Kinect (Microsoft). However, they all allow from 30 FPS (Kinect) to 60FPS (SwissRanger), making possible real-time interaction. While many algorithms exist for analyzing 2D images, few are dedicated to the treatment of three-dimensional point clouds, which represent a large amount of data. Even if this system does not require markers, it remains sensitive to potential problems of occlusion inherent to any system based on a camera.

Any system has drawbacks, of course, but we believe that the use of transparent devices brings a substantial gain for the user compared to current devices. This advantage, associated with the precision of the depth map given by the 3D camera, motivated our choice of such a camera for our system. We chose a SwissRanger camera because it had the best “frame rate / resolution” ratio at the time of the developments and the Kinect was not yet available for sale. Our work can nevertheless easily be transposed to the Microsoft’s Kinect.

## 3. Objectives, general assumptions and working hypotheses

After the state of the art presented above, we propose a modification on the Norman’s model of human-computer interaction [Norman, 1988] (see Figure 1).

In our system we have no device/user or computer/device feedback because the system is completely transparent to the user. There are no sensors or hardware to wear and thus no possible feedback from the device. This is in line with [Fuchs et al., 2003] who claim that ideally the motor responses (in the context of motion capture) must be transmitted without physical medium between man and machine. The system we have developed follows this idea and the Figure 1 (right) presents its functioning. The cycle (in black) on the diagram corresponds to what the user perceives. When the user moves his hand, his action is visually sent back to him by the computer: the visual feedback combined with the absence of hardware to wear make the device totally transparent. The loop (in gray) on our diagram corresponds to what actually happens. The user moves his arms in space, the device (a 3D camera) sends the information to the computer that can determine the position and orientation of the user's hand, and the actions he is doing. Finally, a visual feedback is returned to the user. The absence or negligible of latency between the user’s actions and the visual feedback from the computer is very important. Treatments should be done in real time and with the lowest latency possible.

Our scientific objective is to demonstrate that, by allowing the user to interact directly with the VE, without any intermediary (i.e. actuator), our hand motion capture system is better in terms of performance and acceptability than a device based on the Norman's model (e.g. mouse). We would like our system gives an advantage over other devices, regarding criteria that can possibly reduce the interest of our system (hand size and expertise in VR).

To meet this objective, we conducted a comparative study on two devices (i.e. mouse and 3D Cam) used to perform an objects selection task in a 3D VE. We focused our study on performance (execution speed of the task) and acceptability (perceived usefulness, usability and immersion). To do this, we formulated the following experimental hypothesis:

*Markerless hand motion capture is more efficient and better accepted regarding the selection of objects in a 3D virtual environment than a classic 2D device like a "mouse" regardless of hand size and expertise in VR.*

This hypothesis links 2 dependant variables (performances (DV1) and acceptability (DV2)) to 3 independent variables (device (IV1), size of the hand (IV2) and expertise in VR (IV3)).

To test this main experimental hypothesis, we formulate 5 finest experimental hypotheses in order to isolate each dependent variable and independent variable, which we call operational hypothesis (OH). They are summarized in Table 1.

| OH n° | DV depending on IV | Title |
|---|---|---|
| 1 | DV1 depending on IV1 | Performances (in terms of speed) are better with the 3D Cam than with the mouse |
| 2 | DV2 depending on IV1 | Acceptability (in terms of immersion feeling, usefulness and usability) is better with the 3D Cam than with the mouse. |
| 3 | DV1 depending on IV2 | Performances (in terms of speed) when executing the task with the 3D Cam are not significantly different depending on the size of the user's hand |
| 4 | DV1 depending on IV3 | Performances (in terms of speed) are better for people with good expertise in virtual reality. |
| 5 | DV2 depending on IV3 | Acceptability (in terms of immersion feeling, usefulness and usability) is the same whatever the expertise in virtual reality. |

**Table 1**. Operational hypothesis

## 4. Methodology

### *4.1. Application design*

*4.1.1. Interaction*

The application was designed for the experimentation in order to allow a task of objects selection in a 3D environment. We chose to design a simple virtual environment with the constraints to be able to move the hand (in the virtual environment) at a 1:1 scale, in order to highlight the interest of the hand motion capture without markers, and to obtain results that can be adapted to immersive environments such as “CAVE”. The application is therefore based on the recognition of the hand position (x, y, z coordinates) and the “open hand” and “closed hand” states to match with the real selection gesture.

As our experiments try to compare the manual interaction without markers to a usual interaction device (a mouse), a particular care has been taken to make the two systems equivalent in terms of features and movement amplitude. In both cases the user moves an avatar of the hand in a 3D space. This movement is either indexed on the movement of the hand of the participant (the latency is here negligible) or on the mouse movements. For the “3D Cam” device, we respect the 1:1 scale in order to best comply with the real conditions of use (a move of 1cm of the user’s hand = a move of the hand avatar of 1 cm in the VE). On the contrary, we left the original settings for the mouse in order to preserve the classical paradigm of interaction (i.e. a movement of the mouse induces a greater displacement of the cursor on the screen): the displacement field mouse is therefore restricted to a few centimeters.

When the interaction with the hand allows the movements along the 3 dimensions, it was necessary to adapt the interaction with the mouse. We chose the most common implementation: move in the plane of the screen while moving the mouse and depth moves while moving the mouse (up and down) with the right button pressed (like zooming functionality in some software). For depth moves, the use of the mouse wheel was rejected in view of the

weakness of this system in terms of ergonomics, ease of use and time requested to perform the task. With our system, the selection is made by closing the hand on the object and by left-clicking with the mouse.

*4.1.2. Workspace and tasks*

The application consists of a cubic playing space of 30cm edge (30cm x 30cm x 30cm). The virtual environment is a cube with one face removed allowing the user to see the objects to touch, positioned inside the cube (see Figure 5). The task consists of reaching 10 small boxes of different colors numbered from 1 to 10, one after the other in ascending order. Once the user is well positioned on the correct box, he has to make it disappear, either by a click (mouse) or by closing the hand (3D Cam). It can then proceed to the next box. The participant has to perform this task with the two systems.

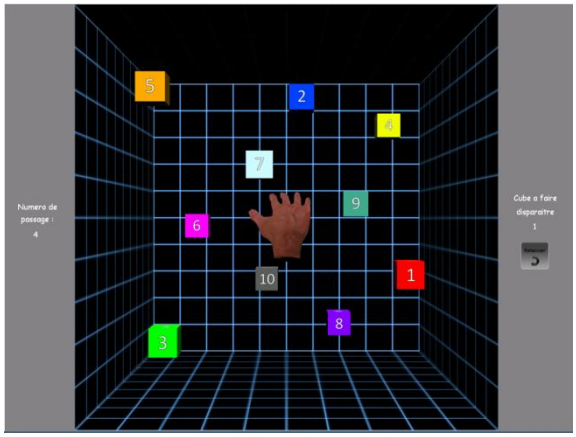



**Figure 5.** Virtual environment – playing space

The game is displayed in perspective and the user views the virtual environment in stereo (active). He can thus estimate in a finer and more precise way the relative distances between each of the boxes. The positions of the 10 boxes are fixed for each system but change between the mouse and the 3D Cam, so that the user cannot memorize the positions of the targets, avoiding thereby any bias due to memorization effect.

*4.1.3. Apparatus and equipment*

The user sits at a table, a 23" screen facing him. The playing space is in front of him, between the table and the 3D Cam located 80 centimeters above (Figure 6).To acquire the 3D position and the opened or closed states of the user's hand, our system is based on a 3D camera (Mesa Imaging a SwissRanger SR4000)[1].

[1] http://www.mesa-imaging.ch/

**Figure 6.** *A participant during the experiment with the mouse (left) and the 3D Cam (right)*

### *4.2. Experimental protocol*

This section describes the empirical study conducted to evaluate the performance (speed of execution of the task) and the acceptability (perceived usefulness, usability and immersion) of the markerless motion capture system we have developed.

#### *4.2.1. Participants*

This study involved 71 participants (21 women and 50 men) aged between 19 and 57 years (average age=27.7 years, SD=8.5). All were experienced users of computers and had a bachelor degree level. They were (for the majority) students for a master degree in virtual reality, or members of a Virtual Reality Laboratory (PhDs, Research Engineers). Participants were characterized by the size of their hand and their degree of expertise in Virtual Reality. These measurements are discussed in the “Measures” subsection.

#### *4.2.2. Procedure*

After participants filled in the identification questionnaire, the experiment and the challenges thereof have been presented to them. A one page written explanation was also given to each participant, as a reminder. The experiment begins with a training session. During this session, limited to just one minute, participants must select 10 times consecutively a randomly positioned cube in space, with each device (mouse and 3D Cam). Participants could then perform the experimental task. Experiment time was limited to 5 minutes to avoid a feeling of despondency by the participants. The elapsed time was not displayed on the screen to avoid stressing the participants. The only visible indication was the number of the current box to select.

#### *4.2.3. Collected data*

Participants' interactions with both devices were recorded throughout the whole experimentation. All the events, whether generated by the participants (moving of the hand, selection, validation ...) or by the system (beginning and end of the experiment, different times, etc.) were automatically marked, dated and saved. Before the experiment, participants were instructed to complete an identification questionnaire (a PDF form). This first questionnaire allowed us to establish the level of expertise in virtual reality of each participant. Subjective judgments and preferences were elicited from post-experimentation questionnaires (a PDF form). Participants were asked to rate various characteristics of the “3D Cam” system and the mouse on criteria like usefulness, usability or immersion. They could also evaluate the task and suggest improvements on the proposed system. Questionnaires contained open questions,

"yes-no" questions as well as Likert scales, where participants were asked to indicate their adequacy degree in relation to a particular criterion (Likert, 1932).

*4.2.4. Measures*

We want to compare the interaction with the 3D Cam and the interaction with the mouse in terms of performance and acceptability according 2 criteria: hand size and expertise in VR. First, we will explain the method used to calculate the expertise in virtual reality and then how we measured the size of the hand. In a second step, we will present the measures of performance and acceptability.

4.2.4.1. Expertise in Virtual Reality

The expertise was calculated from the participants' answers to certain questions of the identification questionnaire. All these questions were Likert scales with 5 modalities. Participants must then position themselves from "not at all" to "excellent knowledge". All the questions were grouped into three themes: general knowledge in VR, use of devices related to VR and use of specific software related to VR. Each modality has been converted into a score: 1 = lowest level, 5 = highest level. Scores were summed for each participant, according to different weights. Three groups of level of expertise were then established (Table 2).

| **Level of expertise** | Beginner | Average | Expert |
|---|---|---|---|
| **Score** | From **0** to **1,5** | From **1,5** to **3** | From **3** to **5** |
| **Workforce by group** | 14 | 33 | 24 |

**Table 2.** *Level of expertise based on the score calculated from the responses to Likert scales*

4.2.4.2. Hand size

We obtained the size of the hand of each participant by measuring the distance, fingers outspread, between the extremity of the middle finger and the extremity of the wrist just below the palm. Figure 7 illustrates the method of measurement.

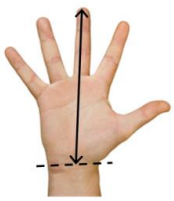


**Figure 7.** *Method used to measure the hand size*

To divide the participants into 3 groups, we calculated the median (18.55) and the quartiles (17.50 and 19.30). The measure of a medium size hand is contained between 17.5cm and 19.3cm. Thus, a "medium size" hand has its measure between 17.5 cm and 19.3 cm. Moreover we observe that the median value calculated on all of our participants is very close to the average values that can be found in the literature (e.g., Ilayperuma et al., 2009). The distribution of the participants is presented in Table 3.

| **Size of the hand** | Small | Medium | Big |
|---|---|---|---|
| **Workforce by group** | 18 | 35 | 15 |

**Table 3.** *Size of the hand in function of the calculated score from the responses to the Likert scales*

4.2.4.3. Performances

The total execution time of the T task (in seconds) served as a basis for studying the performance of the participants. This is the time needed to select the 10 numbered cubes. The execution time begins when the participant confirmed orally that he is ready to begin the experimentation (after the training session).

4.2.4.4. Acceptability

The study of acceptability is based on the participants' answers to the post-experimentation questionnaire:

- "Yes-no" questions, for which participants had to show preference for one or the other device, for example: *What system, mouse or 3D Cam, did you find easier to use?*
- Likert scales with six modalities in which participants must give their adequacy degree in relation to a particular criterion (from *no interest* to *great interest*), for example: *How would you rate the interest of 3D Cam for real time interaction?*

We studied the *acceptabili*ty r*egarding three* criteria: useful*ness, immersion and usability. As stated above, this last criterion w*ill be evaluated according to effectiveness, efficiency and satisfaction.

*4.2.5. Validity*

We implemented different techniques to ensure an internal and external validity of this experimentation. We tried to minimize the influence of certain variables that it is difficult to fully control, such as social background of the participants. The parasitical variables from the physical environment (like: lighting, noise, temperature) were controlled, because the experiment took place in the same room, with air conditioning and no windows.

Given the relatively small number of participants (and especially women), it seemed risky to constitute two separate groups to evaluate each system. Thus, each participant alternately evaluated each system. Pairing allows us to control interference factors, because they are constant throughout the experiment, but may induce the disadvantage of introducing uncontrolled learning effects (i.e. each participant acquires, during the experiment, a certain level of practice, and can accomplish tasks more quickly). The order of execution may therefore play a significant role on performances and preferences (motivation and fatigue are not constant from one subject to another). We defined two groups of subjects, G1 and G2, to counterbalance the execution order. Thus, subjects in G1 began the experimentation by performing the task with the mouse, while subjects in group G2 started experimenting with the 3D Cam. Each group includes the same number of men and women (2 groups of 10 women and 25 men). The assignment of participants to a group or the other (G1 or G2) was made randomly; in order to not involuntary promote one or the other system. The position of the different boxes was predetermined and was the same for each participant, but different between the two devices.

## 5. Results

### *5.1 Participants' performance results*

#### *5.1.1. Statistical analyses and preliminary tests*

First, we conducted tests of normality (Kolgomorov-Smirnov) on the total execution time of the task for each device, to assess the type of statistical test to conduct afterwards. As the distributions were not normal, we used transformations; although their use is debatable (Pallant, 2007), they are suggested in many statistics textbooks and implemented by many authors. In addition, nonparametric tests are limited for studying the interaction effects on mixed models. When appropriate, a qualitative analysis of the shape of the distributions was used to determine the mathematical transformation to apply. In fact, all the tests done for the performances belong to the parametric tests family. To analyze the influence of the device type on participants' performances, we used a Student t-test for paired samples. To analyze the influence of hand size and experience in virtual reality on performance in relation to the device, we performed a mixed-design ANOVA composed of an intra-subject variable (the device) and an inter-subject variable (size of the hand or experience in VR). When needed, analysis of the simple effects was performed using a one-way ANOVA. For the eventual analysis of the main effects, we conducted post-hoc LSD tests. We used the conventional threshold probability $p < 0.05$ for all these tests.

The execution times measured for the mouse didn't follow a normal distribution (Kolmogorv-Smirnov $Z = 1.680$, $p = 0.007$), contrary to those measured for the 3D Cam. We have therefore decided, based on the shape of the distribution, to apply a logarithmic transformation (Log10).

#### *5.1.2. Influence of the device type on performances*

There is no significant difference in participants' performance between the mouse and the 3D Cam ($t = 1.363$, $p = 0.177$). The results are presented in Table 4**Table** .

| | | Original data (seconds) | | Data (Log10) | |
|---|---|---|---|---|---|
| | | 3D Cam | Mouse | 3D Cam | Mouse |
| Mean | | 45.89 | 49.12 | 1.65 | 1.67 |
| Standart Deviation (SD) | | 13.29 | 19.15 | .12 | .15 |
| t-Test | t | | | 1.363 | |
| | p | | | 0.177 | |

**Table 4.** *Means of the execution time with the 3D Cam and the mouse - Original data and transformed ones (Log10)*

Discussion

These results don't confirm OH1 (see Table 5). Although they could seem relatively neutral, are nevertheless encouraging. Besides a weak tendency for the 3D Cam, that we have to confirm or disprove with more participants, we can already say that the system based on 3D Cam has an equivalent level of performance than the mouse for the considered task. Despite the care taken to make the 3D Cam system as simple as possible, the mouse keep its major advantage over any other system, that is its frequent use, which tends to minimize the learning time, even in the case of an original or new task. These results encourage us to realize a study with longer tasks to further highlight the differences between the 2 systems.

*5.1.3. Interaction between the expertise in virtual reality and the device on performances*

We defined three groups of expertise in Virtual Reality: beginner, average and expert. A mixed-design ANOVA (expertise in VR *x* device) didn't show any effect of the expertise or the device, but a significant interaction between these two factors ($F(2, 69) = 4.150$, $p = 0.020$). To study the influence of the simple effects (i.e. the influence of the expertise in VR on the performances for each device) we conducted a one-way ANOVA: persons with an expert or average level in Virtual Reality are significantly more efficient than those qualified as novices, both in using the mouse or the 3D Cam (Table 5). These differences are illustrated in Figure 8.

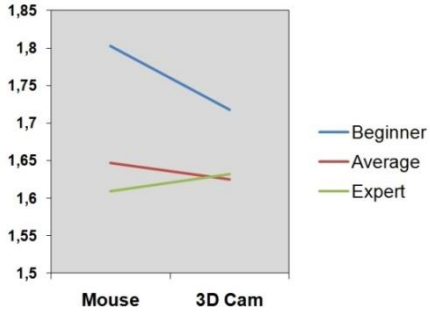



**Figure 8.** *Differences of performances for each device depending on the expertise level in VR*

| | Beginner *vs*. Expert | Beginner *vs*. Average |
|---|---|---|
| Mouse | $p < 0,005$ | $p < 0,005$ |
| 3D Cam | $p = 0,027$ | $p = 0,012$ |

**Table 5.** *Results of the performances analysis based on the expertise in VR for each device (LSD tests)*

Finally, we compared the measured performances with each device separately for each group of expertise. There is no significant difference between the two devices for the expert and average groups. In contrast, for the beginner group, the performances are significantly better with the 3D Cam ($t = 2.650$, $p = 0.020$).

Discussion

The different results reported here provide several answers to our questions. On the one hand, we can see that, in accordance with OH4, the expertise in virtual reality has an impact on performances for our task (similar effect with the 2 proposed devices). On the other hand, we note that the 3D Cam provides a significant advantage for participants qualified as beginners. The immediate takeover of our system, without resorting to an interaction metaphor, may explain this result. People with a greater expertise have already met the problem of 3D positioning with a simple device (keyboard + mouse or mouse-only). New devices or visualizations can lead to rejection, if they do not provide an immediate benefit to the user. The last result is encouraging since it confirms the value of our system especially for beginners.

*5.1.4. Interaction between hand size and device on the performances*

A "hand size *x* device" ANOVA analysis does not indicate any interaction between the size of the hand and the device ($F(2, 65) = 1.288$, NS). There is, however, a main effect of the hand

size on the performance (F (2, 65) = 3.937, p = 0.024). LSD tests indicate a significant difference in performances between participants with small hands and participants with big hands, in favor of the latter (R = 0.1117, p = 0.007).

| | | Small | Medium | Big |
|---|---|---|---|---|
| Mouse | Mean | 1.74 | 1.66 | 1.60 |
| | Std. dev. | 0.17 | 0.14 | 0.07 |
| 3D Cam | Mean | 1.68 | 1.65 | 1.60 |
| | Std. dev. | 0.16 | 0.10 | 0.07 |

**Table 6.** *Performances based on the size of the hand for each device*

Discussion

People with small hands are less efficient. Since there is the same effect for the mouse, it is difficult to give any explanation or even formulate a hypothesis. This result is however in disagreement with OH3, because the size of the hand has an influence on the performances with our system. Nevertheless, the fact that these results are similar with the mouse forces us to remain cautious about this conclusion and requires to conduct additional tests to provide an explanation. In any case, it seems obvious that the type of device is not directly responsible for these differences.

*5.1.5. Synthesis on the performances*

Our study on the performances was conducted in several stages, related with our assumptions. First, we studied with which devices participants got best performances, then we conducted a more detailed analysis to measure the impact of two factors on the performances: expertise in virtual reality and size of the hand.

Contrary to OH1, the 3D Cam device is not superior to the mouse for the selection task in a 3D virtual environment. Nevertheless, the performances obtained with the 3D Cam are equivalent to those obtained with the mouse which is, given this new type of interaction, very encouraging. The results are mixed for the two considered factors: we observe similar results for the 2 devices, with better performances for people with hands characterized as “big” (contrary to OH3) and for people qualified as “experts” or “average” in Virtual Reality (in accordance with OH4). The fact that we obtain similar results for the two devices appears to limit the role of these two factors.

It would be interesting to study on one hand the evolution of performances over a longer period to see if a longer use can lead to a superiority of the 3D Cam over the mouse, which benefits of a much bigger practice. For a selection task, the 3D Cam device can already replace the mouse, without loss of performances. Now it remains to study the acceptability of this system compared to the mouse. By studying both performances and acceptability, we will be able to establish the interests of the 3D Cam and, more generally, of the markerless motion capture for simple tasks such as objects selection in virtual environments.

### *5.2 Acceptability*

*5.2.1. Statistical analyses*

As it has been previously defined, acceptability was studied in terms of utility, immersion and usability; this latter is composed of the effectiveness, the efficiency and the satisfaction.

For yes-no questions, confidence intervals were calculated, in order to make an inference on the observed proportions of participants who chose the answer 3D Cam, to be consistent with the research hypotheses. We use the Wilson's interval method, also called the score interval (Wilson, 1927), which is considered more efficient than the Wald's interval (Gagnon, 2006). In the case of Likert scales, we assigned a score for each modality, and we compared the means between the 3D Cam and mouse devices with the signed rank Wilcoxon test. The calculated scores were also qualitatively illustrated by example from post-experimentation questionnaire.

To study the effect of the expertise in virtual reality on the acceptability, two kinds of tests were used. For yes-no questions, we applied the Pearson's Chi-square test; when the theoretical numbers were less than 5 or when factors had two modalities, we used Fisher's exact test. For the Likert scales allowing the comparison of the two devices, we made a mixed-design analysis of variance (ANOVA), although the distribution of our measures of performances deviates from normality. However, we rely on the observation of Winer (Winer, 1971) on the robustness of the ANOVA with the type 1 errors.

First, we compare perceived acceptability by the participants for each device. We then study the influence of the expertise in virtual reality on the acceptability.

*5.2.2. Influence of the device on the acceptability*

Proportions in favor of the 3D Cam and the mouse and responses to Likert scales are summarized in Table 7. They will then be described and discussed in the following sections.

| | | | **Proportion of the participants in favor of** | | **Mean Likert scale** |
|---|---|---|---|---|---|
| **Criterion** | | **Question** | **3D Cam** | **Mouse** | |
| **Utility** | | Device perceived as the most appropriate for the targets selection task | **85,7 %** | 14,3 % | |
| | | The 3D Cam is perceived as bringing a contribution compared to the mouse | | | 5,06 |
| **Usability** | **Effectiveness** | Device perceived as the most effective in terms of positioning speed | **75,7 %** | 24,3 % | |
| | | Device perceived as the most effective in terms of positioning accuracy | 41,4 % | **58,6 %** | |
| | | Device perceived as the most effective generally | **66,2 %** | 33,8 % | |
| | **Satisfaction** | Device perceived as the more comfortable | 32,4 % | **67,6 %** | |
| | | Device that the participants feel the more comfortable | 47,8 % | 52,3 % | |
| | **Efficiency** | Perceived fatigue with the mouse | | | 1,32 |
| | | Perceived fatigue with the 3D Cam | | | 2,58 |
| | | Device perceived as the easiest to use | **59,2 %** | 40,8 % | |
| **Immersion** | | Device perceived as the most immersive | **98,6 %** | 1,4 % | |

**Table 7.** *Proportions in favor of the 3D Cam and the mouse and responses to Likert scales depending on each criterion*

5.2.2.1 Perception of utility

The 3D Cam device is generally perceived as more useful than the mouse for the task. On one hand, it is more appropriate for the selection of targets, according to the computed confidence interval (75.6% - 92%). The collected data allow us to qualitatively justify this observation; the 3D Cam is considered "easier and faster" and allows a "movement in perfect consistency with the task to do". Indeed, to "grasp" the object, the 'grasp with the hand' scheme is the most natural and realistic". On the other hand, the 3D Cam is considered as bringing a high contribution compared to the mouse (mean score = 5.06 - SD = 0.886). The analysis of the scores distribution (Figure 9) also shows that the majority of the participants rated the contribution as "high" (30/70 or 42%) and "very high" (24/70 or 34%).

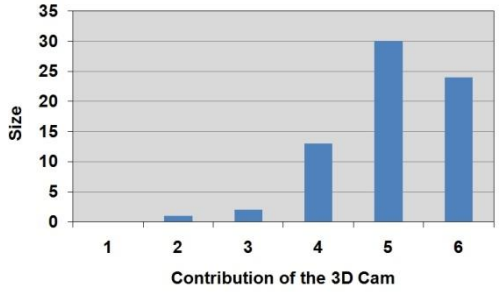



**Figure 9.** *Distribution of the contribution of the 3D Cam compared to the mouse*

5.2.2.2 Perception of usability

Usability is studied in terms of effectiveness, efficiency and satisfaction.

The 3D Cam device is perceived more effective than the mouse for rapid positioning (confidence interval 64.5% - 84.2%) and for overall efficiency (confidence interval 54.4% - 76, 3%). Concerning the accuracy of positioning, it is not possible to assess the superiority of one or the other device (confidence interval 30.6% - 53.1% for the 3D Cam). Participants justify the superiority of the 3D Cam for positioning quickness by "the response in real time of the movements changes [which allows to] more easily manage the movements of the hand" and by "an easier and faster displacement in depth than with the mouse". For the overall efficiency, it is because from "a technical standpoint, the fact that the 3D Cam can track our movements in both plan and depth makes the task easier and maximizes the effectiveness of this one".

The 3D Cam device is not perceived as being more satisfying than the mouse. Indeed, in terms of comfort, the 3D Cam is judged inferior to the mouse (confidence interval 22.7% - 43.9%). For the comfort feeling, the participants are mixed (confidence interval 36.4% - 59.4% with the 3D Cam). If the 3D Cam is considered as "intuitive, fast, efficient, natural, realistic, immersive, fun, direct and unmediated", participants find "more tiring to keep her arm tensed in the air than to put it on the table".

Finally, the superiority of the 3D Cam in terms of efficiency is only partially established. The 3D Cam is mainly seen as being easier to use than the mouse (confidence interval 47.6% - 69.9%), because "interactions are more natural" and "the movement is much more intuitive than with the mouse". On the contrary, the 3D Cam is perceived as more tiring according to the Wilcoxon test for paired samples performed on the scores obtained with the Likert scales ($Z = 5.486$, $p < 0.005$). This result is confirmed by the scores distribution (see Figure 10).

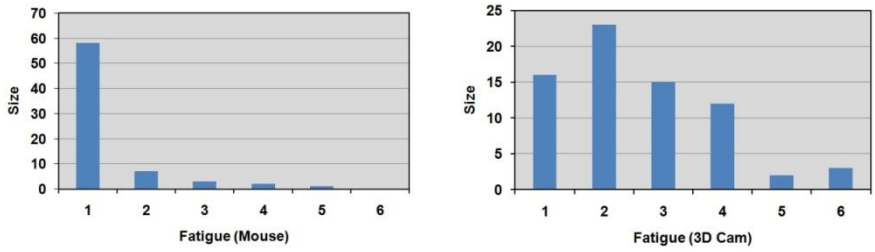



**Figure 10.** *Scores distribution concerning the perceived tiredness with each device ("1" represents an absence of tiredness, "6" a high tiredness)*

5.2.2.3 Perception of immersion

The 3D Cam is generally perceived as being more immersive than the mouse by 70 of the 71 participants in the experiment (confidence interval 92.4% - 99.8%). Participants who judged the 3D Cam more immersive than the mouse justified it by the fact that "the perspective associated with the 3D Cam system gives the sensation of reaching the limit of the scenery when putting the hand out" and by the fact that "having no device helps [the user] to be more focused on the visual representation".

5.2.2.4 Discussion

Overall, the acceptability is better with the 3D Cam which is partially in accordance with OH2. It is undeniable in terms of usefulness and sense of immersion. Concerning the usability, the superiority of the 3D Cam is less obvious, at least on the efficiency and satisfaction criteria. The main reproach is the fatigue and the resulting lack of comfort due to the position of the user, who is seated and must stretch out his arm. Compared with the mouse, the difference is actually very important. This issue is obviously identical with other means of capture, such as data gloves or markers (for IR tracking). It is nevertheless necessary to keep in mind the potential and aimed use of our system in immersive environments like "CAVEs", where the scale is 1:1 and where the user is standing. It is thus easier to adopt a more relaxed posture, with arms resting partially near the bust. The advantage of the greater sense of immersion allowed by our system will take all its sense in this case. Our task, yet simple, of objects selection in a three-dimensional environment, highlights the weaknesses of the mouse for the positioning precision and confirms the interest of the creation or the use of devices suited to tasks in virtual environments.

*5.2.3. Interaction between expertise in Virtual Reality and the device on the acceptability*

There is no significant difference according to the level of expertise in VR in the assessment of the most suitable device for the task of targets selection ($\chi^2 = 0.096$, $p = 0.953$) nor for the judgment of the contribution of the 3D Cam (Kruskal-Wallis H = 1.993, NS).

The participants' expertise in VR doesn't influence the perception of usability. Regarding the effectiveness, there is no significant difference between levels of expertise in VR concerning the most effective device in terms of positioning speed ($\chi^2 = 0.350$, $p = 0.839$), in terms of positioning accuracy ($\chi^2 = 0.651$, $p = 0.722$) and concerning the device considered overall the most effective ($\chi^2 = 1.459$, $p = 0.482$). It is the same for the perception of satisfaction, for comfort ($\chi^2 = 1.312$, $p = 0.519$) and feeling of ease ($\chi^2 = 0.673$, $p = 0.714$). Finally, for efficiency, there is no significant difference according to the level of expertise in the VR on the appreciation of the easier device to use ($\chi^2 = 0.428$, $p = 0.807$). An ANOVA doesn't highlight a significant interaction between the fatigue felt for each device and the expertise in VR of the participants ($F(2.68) = 0.946$, NS).

Whatever the usability, the usefulness and the immersion criterion, there is no significant effects of the expertise in VR on the acceptability; thus these results confirm OH7.

*5.2.3. Synthesis on the acceptability*

We replicated the plan used for the performance analysis that is a comparative study between the 3D Cam and the mouse, then a detailed study of the influence of the expertise in virtual reality on these results. According to the participants, the 3D Cam provides a significant advantage in terms of perceived usefulness and immersion, but remains lower than the mouse in terms of efficiency and satisfaction. A greater feeling of fatigue is the main raison to explain this last result, due to the sitting position with the arm put out. Even if this point is critical, it is more related to the selection task in a 3D environment than to the strict use of the device. It will be necessary to find a solution to allow the arm to have some rest without sacrificing the realism and the immersion. Finally, it should be noticed that fatigue could be reduced for interactions in a CAVE where the user is standing. The 3D Cam can favorably replace in this situation a hand-held device (e.g. ART Flystick).

Expertise in virtual reality does not influence the acceptability in accordance with OH5. In the questionnaires, participants did not report cognitive difficulties in the use of our system; the fact that there is no need to use interaction metaphors is thus a gain. The use of a technique to track the hand without markers doesn't allow to provide haptic feedback. Apart the visual feedback of the position of the hand avatar in space, users do not have other proprioceptive clues. However, participants did not mention the lack of haptic feedback when collisions occurred between the hand and the boxes during the selection. These results demonstrate on one hand the value that our system can bring for tasks on a computer screen where the feeling of immersion is usually low and the position in space can be problematic [Banos et al., 2004] and on the other hand the interest that the porting of our system could have in an immersive and stereoscopic environment at a 1:1 scale.

## 6. Conclusion

The interaction in virtual environments is mainly based on the hand which covers in the real world nearly 70% of motor abilities of man. The rapid democratization of virtual reality tools, and in particular the spreading of 3D stereoscopic displays (e.g. for home entertainment like games or movies), makes it necessary to develop appropriated interaction means suited to real time 3D. Indeed, no device currently meets the criteria of reliability, ease of use, low cost and transparency for the user. This is particularly true for data gloves that are the most used but still combine these problems today. The concept of immersion is also very important in virtual reality and the absence of intermediate metaphor could simplify the interface design (Richir and Fuchs, 2006). If transparency can increase the immersion feeling, it can also be a major advantage when it is not possible to install equipment on users (e.g. for people with autism). Therefore we believe it is now necessary to focus towards a solution that would be simpler to implement, less costly and more easily adopted by users from the general-public. Interest of motion recognition technologies based on the use of cameras without markers seems then obvious. The main obstacle for markerless hands motion capture is that it is not currently efficient enough with the standard cameras technology. But computer image processing techniques, combined with new devices (e.g. 3D cameras like Microsoft Kinect) and sufficient computing power (GPU and CPU) for processing the data stream, are now sufficiently mature to enable real-time interaction with the advantages mentioned above. It remains to prove the benefit of this technology for real-time interaction in virtual environment, and explore its potential.

The aim of this paper was to demonstrate, through a simple selection task in three dimensional virtual environments, the value of such a system compared to a device commonly used for this type of task. For this, we conducted an experiment on 71 participants (50 men, 21 women). The task consisted of successively selecting 10 cubes in a virtual environment, alternately with the mouse and with a 3D camera. We formulated the general hypothesis that the 3D camera (3D Cam) brings a significant advantage in terms of performance and acceptability, regardless of the size of the user's hand and his experience in virtual reality. This assumption guided our choice for quantitative and qualitative analysis. The performance corresponds to the measure of the total execution time of the task with each system. The qualitative analysis is based on the study of the acceptability and, more specifically, the usefulness, the usability and the feeling of immersion; these data were collected through a post-experimentation questionnaire. Different statistical tests were conducted to ensure validity and reproducibility of the results.

Performances with the 3D Cam are not significantly better than those obtained with the mouse. This result has to be put into perspective with the very short time of experimentation, which is a disadvantage for the 3D Cam given that the mouse receives a higher level of practice, despite the learning phase at the beginning of the experiment. This first result does not confirm our hypothesis, but also doesn't refute it. Virtual reality and the size of the hand do not influence the performances obtained with the 3D Cam. Our results indicate that average and experts participants have better performances than beginners. Similarly, people with medium and big hands are faster to complete the task than people with small hands. The perception of acceptability is divided: even if the 3D Cam is judged more useful and provides a better sense of immersion, participants criticize the usability, because of a higher feeling of fatigue with the 3D Cam. The position of the user, sitting with the arm put out without support justifies this decision. Expertise in virtual reality does not influence the perceived acceptability by participants. Taken together, these quantitative and qualitative results clearly demonstrate the interest of a markerless device for selection tasks in virtual environments. Indeed, if performances are equivalent, the sense of immersion is higher with our system. Interaction, with no intermediary metaphor, provides an immediate understanding to the user and a great ease of use. These different results reveal a paradox: if a longer use would reduce the bias of the learning period and would positively impact on performances, it would degrade at the same time the perceived usability and more specifically would cause more fatigue (which is already important for a less than one minute use). If the interest of our system is proven, it would be more appropriate in two use case scenarios allowing to reduce the perceived fatigue. On one hand in a context of a standing use such as in an immersive environment with a 3D stereoscopic vision: the user would not have to work with the extended arm and could rest his arm on the upper body. On the other hand, in the context of seated use, for occasional gestures, such as in the case of controlling remote devices (home automation).

Our general hypothesis is partially verified. Some results are nevertheless difficult to explain: it is curious to see a difference of performances relatively to the size of the hand of the participants. In fact, users with big hands, leaving partially the playing space, could cause errors in the position detection of the hand and thus a longer execution time. However, performance is lower with people with small hands. We have to further investigate this result. The best performances obtained by users familiar with virtual reality applications can be explained by a greater propensity of this type of users to adapt to new ways of interaction. It should be noted that, for novices, the results are better with the 3D Cam than with the mouse. This result is interesting in the context of a universal access to new technologies.

After this first study, based on a simple task, we can say that the markerless motion capture of the hand is promising because it offers undeniable advantages in the context of interaction in virtual environments, where realism [Witmer et al., 1998] and the felt presence [Sheridan,

1996] is an important driver for the feeling of immersion. The risk remains that the current enthusiasm for time-of-flight cameras, reinforced by the success of the Microsoft's Kinect, is just a fad. If it does not provide a real added value to the user, the system will not be useful [Loup-Escande et al., 2011] and will be quickly abandoned in favor of other technologies. This work is a first step towards the scientific demonstration of the contributions of this technology. It is a starting point and thus provides a justification for further research on other types of 3D and real time activities, with longer experimentation time, and comparison with more advanced devices.

## 7. Perspectives

We have shown through our state of the art that many devices based on various technologies exist to translate the movements of the hands and the fingers in a virtual environment or simply to recognize gestures in a real environment. However, no “universal” device has been yet developed for manual interaction in three dimensions; the universality can be in this case described by a wide adoption, motivated by a simple and not restrictive use, with high performances, for a relatively low cost. The launching of 3D cameras, called “time of flight” cam, is a first strong signal in this direction. But beyond the technical potential, few research studies have been focused on the real contribution of this technology which is still emerging.

The study reported here, through a simple selection task, demonstrates the interest of this type of camera for real time motion capture. Reproduce the hand movements reliably and accurately is a crucial issue, and our results show that this type of device should already be adopted in the field of virtual reality, because of its superiority in terms of the perception of immersion. There are many perspectives opened by our work, and more generally by markerless motion capture.

This first experiment was based on a simple language (open hand, closed hand) as well as the only consideration of the hand position in a three-dimensional Cartesian coordinates system. We must now extend the capabilities of our system to determine the orientation of the hand and then rebuild the skeleton of the hand in order to consider fine finger movements. The necessary development for the first step is now completed and we have to develop more complex selecting and manipulating task of 3D objects with bimanual interactions in real time. We will be thus closer to real tasks and able to assess the qualitative and quantitative interest of our markerless capture system compared to conventional systems, that is to say, data gloves and optical cameras. For the second step (the skeleton reconstruction), we must be able to accurately identify the movements of the fingers. Further options such as Markov chains or artificial neural networks are alternatives whose interest in the area of recognition is demonstrated. However, even if these techniques are used for character, patterns or shapes recognition, they still are not widely used for motion recognition. But the possibilities seem a priori promising.

The release of the Microsoft “Kinect” in November 2010, based on a quite similar technology to the camera we used in our system, is encouraging in view of the wide dissemination of this type of device, because of a price / performance ratio extremely favorable. The playful scope was very quickly exceeded and the possibilities of this device led to the birth of a large community of developers. It also decided Microsoft to distribute a SDK for Windows because most drivers and source codes were written by open-source community. Works are still poorly organized, focus mainly on the motion capture of the entire body at the expense of the required accuracy for manual interaction, and often remain in a simple demonstrator state.

But the importance of the work done will undoubtedly facilitate the emergence of new uses and the wide adoption of this type of device, if indeed we can attest the interest in various fields.

Today the application areas covered are numerous. Uses can range from health and particularly rehabilitation (Zhou and Hu, 2008; Movea, 2009), manipulation of objects in virtual environments (Schlattmann and Klein, 2009), to the control of systems by movements, for example in the case of home automation. In the field of disability, this kind of system could facilitate interaction of children and adults with multiple disabilities who have difficulty carrying out voluntary movements of the upper limbs. Current activities, only based on switches [Lancioni et al., 2011], could be advantageously replaced by a markerless motion tracking system, with a relatively simple and adapted language to each child. These are just two examples among others and it is now appropriate to explore the range of possibilities.

Scientific, technological and applicative prospects can therefore enable significant societal impact. They justify the continuation of efforts in the field of markerless hand motion recognition.

## 8. Bibliography

Arnaldi B., Tisseau J., Berthoz A., Burkhardt J.-M., B., Coquillart S., Guitton P. et al., "L'interfaçage, l'immersion et l'interaction", In *Traité de la réalité virtuelle*, 3 ed., Vol. 2, p. 524, 2003, Les Presses de l'École des Mines de Paris.

Bach C., Scapin D., "Ergonomic criteria adapted to human virtual environment interaction", *Proceedings of the 15th French-speaking conference on human-computer interaction, Caen, France, 24-31*, 2003.

Banos R. M., Botella C., Alcaniz M., Liano V., Guerrero B., Rey B., "Immersion and Emotion: Their Impact on the Sense of Presence", CyberPsychology Behavior, vol. 7, n° 6, 2004, p. 734-741.

Beaudouin-Lafon M., "Designing interaction, not interfaces", *Proceedings of the working conference on Advanced Visual Interfaces, Gallipoli, Italy, 15-22,* 2004.

Bowman D. A., "Interaction techniques for common tasks in immersive virtual environments: design, evaluation, and application", Thesis, Georgia Institute of Technology, 1998.

Bowman D. A., McMahan R. P., "Virtual Reality: how much immersion is enough?", *IEEE Computer*, vol. 40, n° 7 , 2007, p. 36-43.

Bregler C., Malik J., "Tracking People with Twists and Exponential Maps", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998.

Buchholz B., Armstrong T.J., "A kinematic model of the human hand to evaluate its prehensile capabilities", *Journal of Biomechanics*, vol. 25, n°2, 1992, p. 149-162.

Collet C., "Capture et suivi du regard par un système de vision dans le contexte de la communication homme-machine", Thesis, Ecole Nationale Supérieure de Cachan, 1999.

Coquillart S., Fuchs P., GrosJean J., Hachet M., Bechmann D., Stenberger L., "Les techniques d'interaction pour les primitives comportementales virtuelles", In *Traité de la réalité virtuelle : Volume 2, L'interfaçage, l'immersion et l'interaction,* 3 ed., vol. 2, p. 332, 2003, Les Presses de l'École des Mines de Paris.

Delamarre Q., Faugeras O., "3D Articulated Models and Multi-View Tracking with Silhouettes", *In proceedings of International Conference on Computer Vision*, IEEE Computer Society, Kerkyra , Grece, 1999, p. 716-721.

Di Stefano L., Marchionni M., Mattoccia S., "A fast area-based stereo matching algorithm", *International Conference on Vision Interface*, vol. 22, n° 12, 2004, p. 983-1005.

Dvorak J.C., *The San Francisco Examiner*, 1984, February 19.

Fuchs P., Burkhardt J.-M., Lourdeaux D., "Approche théorique et pragmatique de la réalité virtuelle", *In Traité de la réalité virtuelle : Volume 2, L'interfaçage, l'immersion et l'interaction,* 3 ed., vol. 2, p. 23, 2003, Les Presses de l'École des Mines de Paris.

Fuchs P., Mathieu H., "Les interfaces spécifiques de la localisation corporelle - Introduction", *In Traité de la réalité virtuelle : Volume 2, L'interfaçage, l'immersion et l'interaction,* 3 ed., vol. 2, p. 93-94, 2003, Les Presses de l'École des Mines de Paris.

Gagnon P., "Intervalles de confiance pour une différence de deux proportions", Master thesis, Université de Laval, Laval, 2006.

Geebelen G., Maesen S., Cuypers T., Bekaert P., "Real-Time Hand Tracking with a Colored Glove", *In Proceedings of 3D Stereo Media*, Luik, Belgium, 2010.

Gosselin F., Andriot C., Fuchs P. (2003). "Les dispositifs matériels des interfaces à retour d'effort", *In Traité de la réalité virtuelle : Volume 2, L'interfaçage, l'immersion et l'interaction,* 3 ed., vol. 2, p. 135, 2003, Les Presses de l'École des Mines de Paris.

Gumpp T., Azad P., Welke K., Oztop E., Dillmann R., Cheng G., "Unconstrained Real-time Markerless Hand Tracking for Humanoid Interaction", *In proceedings of 6th IEEE-RAS International Conference on Humanoid Robots*, IEEE Press, Genova, Swiss, 2006, p. 88-93.

Häger-Ross C., Schieber M.H., "Quantifying the Independence of Human Finger Movements: Comparisons of Digits, Hands, and Movement Frequencies", *The Journal of neuroscience,* vol. 20, n° 22, 2000, p. 8542-8550.

Hand C., "A Survey of 3D Interaction Techniques", *Computer Graphics Forum,* vol. 16, n° 5, 1997, p. 269-281.

Hassanpour R., Shahbahrami A., Wong S., "Adaptive Gaussian Mixture Model for Skin Color Segmentation", *International Journal of Computer and Information Science and Engineering*, vol. 2, 2008.

Hayward V., "Toward a Seven Axis Haptic Display", *Proceedings of International Conference on Intelligent Robots and Systems, Pittsburgh, Pennsylvanie, USA*, 1995, p. 133-139.

Herda L., Fua P., Plänkers R., Boulic R., Thalmann, D., "Using skeleton-based tracking to increase the reliability of optical motion capture", *Human Movement Science Journal,* vol. 20, n° 3, 2001, p. 313-341.

Ilayperuma I., Nanayakkara G., Palahepitiya, N., "Prediction of personal stature based on the hand length", *Galle Medical Journal,* vol. 14, n° 1, 2009, p. 15-18.

INRIA-Rhône-Alpes, *La plateforme GrImage*, from http://grimage.inrialpes.fr, 2007

Kato M., Xu G., "Occlusion-Free Hand Motion Tracking by Multiple Cameras and Particle Filtering with Prediction", *International Journal of Computer Science and Network*, vol. 6, 2006, p. 58-65.

Kolb A., Barth E., Koch R., Larsen R., "Time-of-Flight Sensors in Computer Graphics", in: M. Pauly, G. Greiner (Eds.) *30th Annual Conference of the European Association for Computer Graphics*, Munich, Germany, 2009, p. 119-134.

Krout M.H., "Autistic gestures: an experimental study in symbolic movement", *Psychological Monographs,* vol. 46, n° 4, 1935, p. 119-120.

Kuch J.J., Huang, T.S., "Vision based hand modeling and tracking for virtual teleconferencing and telecollaboration", *In Proceedings of the Fifth International Conference on Computer Vision*, 1995, p. 666-671.

Lancioni G.E., Singh N.N., O'Reilly M.F., Sigafoos J., Green V., Oliva D., Lang R., "Microswitch and Keyboard-Emulator Technology to Facilitate the Writing Performance of Persons with Extensive Motor Disabilities", *Research in Developmental Disabilities*, vol. 32, n° 2, 2011, p. 576-582.

Lempereur M., “Simulation du mouvement d'entrée dans un véhicule automobile”, Thesis, Université de Valenciennes et du Hainaut-Cambrésis, 2008.

Likert R., “A technique for the measurement of attitudes”, *Archives of Psychology,* vol. 22, n° 140, 1932, p. 1-55.

Lin J., Wu Y., Huang T.S., “Modeling the constraints of human hand motion”, *IEEE Workshop on Human Motion (HUMO'00)*, 2000, p. 121-126.

Loup-Escande E., Burkhardt J.-M., Richir S., Anticiper et Evaluer l’Utilité dans la Conception Ergonomique des Technologies Emergentes : Une Revue, *Le Travail Humain*, 2011, in press.

MacKenzie C.L., Iberall T., *The Grasping hand* (Vol. 104), Amsterdam, Elsevier, 1994.

Meseure P., Kheddar A., “Les outils et les modèles informatiques des environnements virtuels”, *In Traité de la réalité virtuelle,* 3 ed., vol. 3, p. 141, 2006, Les Presses de l’École des Mines de Paris.

Michoud B., Guillou E., Bouakaz S., “Extension de l’espace d’acquisition pour les méthodes de Shape-from-silhouette”, *Paper presented at the COmpression et REpresentation des Signaux Audiovisuels Conference*, 2006.

Mizuuchi I., “A Musculoskeletal Flexible-Spine Humanoid Kotaro Aiming at the Future in 15 Years Time”, *Mobile Robots - Towards New Applications: Advanced Robotics Systems International*, 2006, p. 45-56.

Mohr D., Zachmann G., “Continuous Edge Gradient-based Template Matching for Articulated Object”, *In proceedings of the International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, 2009, p. 519-524.

Movea, *MotionPod™ Technology*, from http://movea.com/healthcare/motion_pod/index.html, 2009

Napier J.R., “The prehensile movements of the human hand”, *The Journal of bone and joint surgery,* vol. 38-B, n° 4, 1956, p. 902-913.

Nedel L.P., Dal Sasso Freitas C. M., Jacob L., Pimenta M. S., “Testing the Use of Egocentric Interactive Techniques in Immersive Virtual Environments”, *In proceedings of International Conference on Human-Computer Interaction, Zurich, Switzerland*, 2003.

Norman D.A., *The psychology of everyday things*, Basic Books, 1998.

Ouhaddi H., Horain P., “Conception et ajustement d’un modèle 3D articulé de la main”, *6èmes Journées de Travail du GT Réalité Virtuelle, Issy-les-Moulineaux, France*, 1998, p. 83-90.

Pamplona V.F., Fernandes L.A.F., Prauchner J., Nedel L.P., Oliveira, M.M., “The Image-Based Data Glove”, *In proceedings of 10th Symposium on Virtual and Augmented Reality*, João Pessoa, Brazil 2008, pp. 204-211.

Pallant J., *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows* (3 ed.), 2007, Open University Press.

Rautaray S.S., Agrawal A., “A Real Time Hand Tracking System for Interactive Applications”, *International Journal of Computer Applications*, vol. 18, 2011, p. 28-33.

Regh J.M., Kanade T., “Visual Tracking of Height DOF Articulated Structures: an Application to Human Hand Tracking”, *In proceedings of 3rd ECCV*, 1994, p. 37-46.

Richir S., Fuchs P., “La méthode I²I : Interaction et Immersion pour l'Innovation”, *Techniques de l'ingénieur. Télécoms,* vol. TEB3, n° TE5910, 2006, p. 1-9.

Rolland R., Yvain E., Christmann O., Loup-Escande E., Richir S., “E-commerce and Web 3D for involving the customer in the design process: the case of a gates 3D configurator”, *In proceedings of Virtual Reality International Conference*, Laval, France, 2012.

Romero J., Kragic D., Kyrki V., Argyros A., “Dynamic Time Warping for Binocular Hand Tracking and Reconstruction”, *In Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Pasadena, CA, USA, 2008, p. 2289 – 2294.

Saxena A., Schulte J., Ng A.Y., "Depth estimation using monocular and stereo cues", *In proceedings of the 20th international joint conference on Artifical intelligence, Hyderabad, India*, 2007, p. 2197-2203.

Scharstein D., Szeliski R., "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *International Journal of Computer Vision,* vol. 47, n° 1-3, 2002, p. 7-42.

Schlattmann M., Klein R., "Efficient Bimanual Symmetric 3D Manipulation for Markerless Hand-Tracking", *In proceedings of the Virtual Reality International Conference, Laval, France*, 2009.

Stefano L.D., Marchionni M., Mattoccia S., "A fast area-based stereo matching algorithm", *Image and Vision Computing*, vol. 22, 2004, p. 983-1005.

StengerB., "Template-Based Hand Pose Recognition Using Multiple Cues", *In proceedings of Computer Vision*, Springer Verlag, Hyderabad, India, 2006, p. 551-560.

Tosas M., "Visual Articulated Hand Tracking for Interactive Surfaces", Thesis, University of Nottingham, 2006.

Wang R.Y., Popovic J., "Real-Time Hand-Tracking with a Color Glove", *ACM Transactions on Graphics*, vol. 28, 2009, p. 1-8.

Wilson E.B., "Probable Inference, the Law of Succession, and Statistical Inference", *Journal of the American Statistical Association*, vol. 22, n° 158, 1927, p. 209-212.

Winer B.J., *Statistical principles in experimental design, 2nd* ed., 1971, McGraw-Hill.

Witmer B.G., Singer M.J., "Measuring presence in virtual environments: A presence questionnaire", *Presence : Teleoperators and Virtual Environments*, vol. 7, n° 3, 1998, p. 225-240.

Zhou H., Hu H., "Human motion tracking for rehabilitation - A survey", *Biomedical Signal Processing and Control*, vol. 3, n° 1, 2008.