Virtual Reality (2002) 6:21–32

A VE Framework to Study Visual Perception and Action

F. Panerai, M. Ehrette, P. Leboucher

Laboratoire de Physiologie de la Perception et de l'Action (LPPA), CNRS/Collège de France, Paris, France

Abstract: In the real world, vision operates in harmony with self-motion yielding the observer to unambiguous perception of the three-dimensional (3D) space. In laboratory conditions, because of technical difficulties, researchers studying 3D perception have often preferred to use the substitute of a stationary observer, somehow neglecting aspects of the action-perception cycle. Recent results in visual psychophysics have proved that self-motion and visual processes interact, leading the moving observer to interpret a 3D virtual scene differently from a stationary observer. In this paper we describe a virtual environment (VE) framework which presents very interesting characteristics for designing experiments in visual perception during action. These characteristics arise in a number of ways from the design of a unique motion capture device. First, its accuracy and the minimal latency in position measurement; second, its ease of use and the adaptability to different display interfaces. Such a VE framework enables the experimenter to recreate stimulation conditions characterised by a degree of sensory coherence typical of the real world. Moreover, because of its accuracy and flexibility, the same device can be used as a measurement tool to perform elementary but essential calibration procedures. The VE framework has been used to conduct two studies which compare the perception of 3D variables of the environment in moving and in stationary observers under monocular vision. The first study concerns the perception of absolute distance, i.e. the distance separating an object and the observer. The second study refers to the perception of the orientation of a surface both in the absence and presence of conflicts between static and dynamic visual cues. In the two cases, the VE framework has enabled the design of optimal experimental conditions, permitting light to be shed on the role of action in 3D visual perception.

Keywords: Action; Motion capture; Virtual environments; 3D visual perception; Visuo-motor coherence

Introduction

In recent years the use of virtual reality (VR) and virtual environments (VE) techniques in fundamental and applied research has increased enormously [1]. In the neuroscience field, for example, these powerful tools enable the recreation of complex spatio-temporal stimulations and the control of characteristics of the stimuli at the desired level of detail. Complex human behaviours are nowadays studied using state-of-the-art graphics and displays, combined with haptic, kinesthetic and acoustic interfaces [2]. These extraordinary tools stimulate new methodological approaches to understanding information processing in the human brain, and from the point of view of experimental psychology and psychophysics, promise new insights to a richer understanding of currently popular theory of sensory perception [3,4]. Studies concerning visuo-motor transformation, 3D perception and navigation abilities, and at a more abstract level, cognitive functions, are the more promising areas of application of these techniques. Ghahramani and Wolpert [5], for example, have used a VE to create a conflict between the visual and the motor space. Using this approach, they have shown that in visuo-motor learning, the brain may employ modular decomposition strategies. Aguirre et al. [6], used VE techniques in combination with functional magnetic resonance imaging to localise the neural substrate of human topographic spatial learning abilities. Bertin et al. [7] used a VE framework to simulate 2D navigation through a 3D space, studying perception of ego-motion from optic flow (but see also [8]). Amorim et al. [9] have employed a desktop interface to address questions related to higher cognitive functions, such as the behavior of VE users confronted with the problem of relating their spatial orientation and a change in the visual perspective within the VE

Depending on the specific research issues, the characteristics required of a VE can change substantially. For example, investigations in visual perception and action require that the experimenter has control over the coherence of different sensory channels activated during a given task. The absence of cross-modal coupling can lead the central nervous system (CNS) to ignore part the of sensory information which does not conform to real world-like situations, or in the worst case, can cause user sickness [10–12]. For those studies addressing the issues of visual perception during movement, it is of primary importance that the relationship between the observer's movement and the dynamics of the visual images falling onto the retinas be reproduced to the highest degree of fidelity. Recent investigations have proved that, although at different levels, several anticipatory mechanisms exist which help the CNS to build predictions of future sensory and motor events, for example in visuo-motor control loops [13,14] but also in pure visual processing [15,16]. In other terms, in order to recreate the exact circumstances in which the brain exploits synergies of sensory modalities (i.e. learned sensori-motor couplings, predictive mechanisms, combinations of visual and movement sensation, of kinaesthetic and proprioceptive information), sensory information should be spatially and temporally consistent [17,18].

In 3D visual perception, much of the past research has been dominated by the presence of static observers (i.e. observers moving their eyes but not the whole head). This choice was partially due to the difficulty in setting up an interactive VR framework, where coherent sensory information could be experienced. The presentation of a sequence of shots of a visual scene has been the most commonly used paradigm. On the other hand, it is not arguable that vision occurs naturally in presence of complex movements of the eye and of the head. To what extent these reflex or voluntary movements and their consequences could impact visual information processing and therefore perception is a key question which can only be addressed in a VE framework which enables us to reproduce correctly coherent sensory information.

In this work, we describe a VE framework purposely designed to study 3D visual perception during action. The intention was not to compare it with other frameworks (for example, those using HMD technologies), but to show that it offers remarkable advantages in order to set up experimental studies in this field. In order to build the framework, a motion capture device was designed and realised. It is versatile, has low latency, and is extremely precise. It can be integrated to different display interfaces: (1) a large screen, which exposes users to moderate resolution, wide-field, full-immersive conditions; and (2) a desktop monitor, which provides small-field, high-resolution conditions. Previous work [19] led to the observation that factors such as the wide field of view, the coherence between sensory and motor information and the absence of annoying weight on the observer's head, are judged by VE users as very important characteristics which contribute to reproduce the natural conditions of visual exploration.

In two studies, we describe the use of the VE framework in large-field and small-field display conditions. The motion capture device is used to compute, in real time, the 3D position and orientation of the head of the observer. This information enables us, in turn, to determine the exact point of view (PoV) of the observer, and generate the instantaneous and correct image perspective. In order to make available the techniques and the methods developed, we present an overview of them. In the section titled The Motion Capture Device, the motion capture device is described; the modelling of the observer point of view is then illustrated together with a method to compute the visual-display transformations and the calibration procedures used to deal with individual user differences are explained. Finally, the experimental contexts of two studies on 3D visual perception during action performed within the VE framework are introduced. In the first study, the observer explores a virtual scene in large-field, fully immersive conditions, and his/her task is to recover the absolute distance of objects presented in subsequent trials. In the second study, the observer explores virtual planar objects in small-field desktop conditions and his/her task is to estimate the 3D surface orientation.

The Motion Capture Device

The motion capture device (tracker) is a multi-link mechanical structure. It is composed of two sets of rotational joints, the *base* and the *end part*, which are interconnected by a pole. The pole, made out of carbon-fibre, can slide through the base and acts as a translational joint. It is extremely light-weight, non-deformable and can be extended to a maximum length

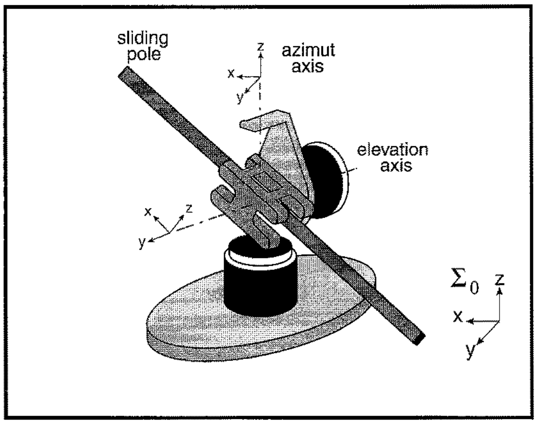

Fig. 1. The base of the motion capture device. Two rotational axes, the azimuth and the elevation, enable the end part of the pole to move linearly (3 DOF) through a working space characterized by a rotational symmetry with respect to the vertical azimuth axis.

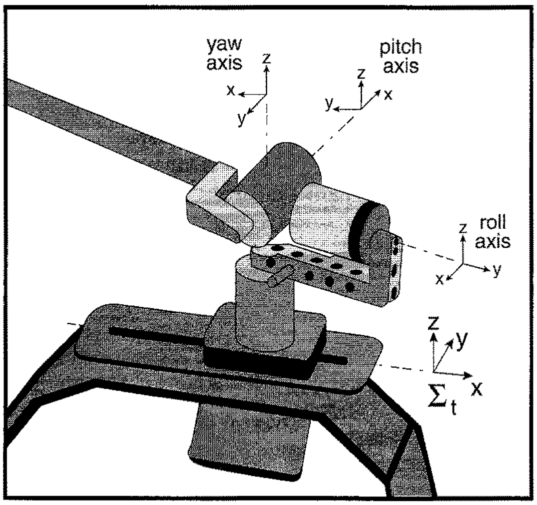

Fig. 2. The end part of the motion capture device. Attached to one end of the sliding pole, a mechanical part composed of three rotational axes, the roll, the pitch and the yaw, enable full rotational movements of the helmet in 3D space (6 DOF).

of 1.5 m. The two rotational axes of the base are positioned perpendicularly (see Fig. 1). They provide the *azimuth* and the *elevation* degrees of freedom of the pole. The figure also shows a mechanical part rigidly fixed to the elevation axis, which contains roller bearings and has two functions: it holds the pole and enables its sliding. The end part of the device (Fig. 2) is composed of three rotational axes, which are arranged in a *pitch-roll-yaw* configuration. The pitch axis is rigidly fixed to the end point of the pole, the yaw axis is fastened to a light-weight helmet. Once the subject wears the helmet, the tracker can determine the position and the orientation (six degrees of freedom, DOF) of his/her head with respect to the base, by means of a mathematical model.

Table 1. Characteristics of the motion capture device. Degrees of freedom (DOF), sensor technology and resolution.

DOF	Sensor technology	Resolution
Azimuth	Optical	$9 \cdot 10^{-3}$ deg
Elevation	Optical	$9 \cdot 10^{-3}$ deg
Translation	Magnetic	10^{-5} m
Pitch	Optical	$36 \cdot 10^{-3}$ deg
Roll	Optical	$36 \cdot 10^{-3}$ deg
Yaw	Optical	$36 \cdot 10^{-3}$ deg

Sensors

Two types of sensor technologies are used. The angular positions of the rotational joints are read through optical encoders. At the base, the encoders have a resolution of $40 \cdot 10^3$ steps per revolution (i.e. $9 \cdot 10^{-3}$ degrees). In the end part, the encoders are less accurate, but also less cumbersome, and have a resolution is of $10 \cdot 10^3$ steps/revolution (i.e. $36 \cdot 10^{-3}$ degrees). The linear displacement of the sliding pole is measured by a magnetic transducer, which has a resolution of 10^{-2} mm. Table 1 summarises this technical data.

Data Acquisition

The linear and the angular encoders output position information in digital format. The position data is acquired using a National Instruments digital I/O board. Latency in acquisition time (T_a) for the six encoders is less than 500 µs on a standard Pentium II processor. The time interval (T_c) to compute the orientation and the position information of the head in 3D space, is dependent on the available processor. In the current configuration (a Pentium II processor), the sum of the two time intervals (T_a+T_c) is less than 1 ms. Therefore, theoretically, the maximum sampling rate can be as high as 1 kHz.

Modelling the Point of View of the Moving Observer

In order to measure the position and the orientation of the observer's head in 3D space, a mathematical

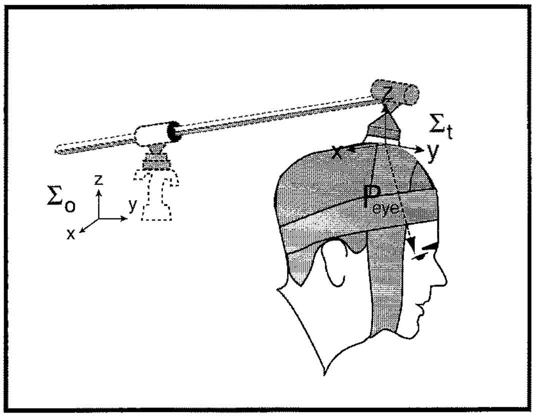

Fig. 3. Modelling of the eye position. The position of the centre of the eye can be described in the helmet reference frame (Σ_t) as a constant vector (${}^{t}P_{eye}$). The same position can be represented by a vector (${}^{o}P_{eye}$) defined in the base reference frame (Σ_o).

model of the multi-link device has been developed. It is formulated on the basis of standard mathematical formalism used for kinematics of robot manipulators [20]. In previous work, the procedure used to derive the model has been fully described and the accuracy obtained in measuring position (in the order of 0.6 mm, worst case) was assessed [21]. The device provides head position information in the form of three Cartesian coordinates and of a rotation matrix which describes the orientation of the helmet with respect to the base of the device. Since we will refer to this mathematical formalism throughout the paper, part of it is briefly reviewed in the appendix. In the next section, we address the problem of how to extend the mathematical formalism to include the modelling of the observer's point of view (PoV), i.e. the position in 3D space of the centre of the eye of the observer.

Head Position Information

The position and orientation of the helmet with respect to the base is completely specified once the angular and linear readings of the encoders, i.e. the *joint variables*, are input to the mathematical model. In the model two different reference frames are associated with the base and the helmet. These frames are respectively Σ_o and Σ_t (see Fig. 3). The model transforms the coordinates of any 3D point relative to the helmet frame into coordinates relative to the base frame. Such a transformation of coordinates can be represented by a 4×4 homogeneous matrix, described as follows:

$$ {}^{o}_{t}T(\underline{q}) \qquad (1) $$

where the q term is the vector representing the six joint variables, q_i (i=1 ... 6). The joint variables are respectively, the *elevation* and *azimut* angles of the pole (shown in Fig. 1), the *translation* of the pole and the *roll-pitch-yaw* angles of the helmet (see Fig. 2). As already mentioned, the mathematical model, represented by Eq. (1), is suitable to describe the position of any point relative to the helmet (Σ_t) and transform its coordinates in the base reference frame (Σ_o). For our purpose, we are interested in defining the 3D location of the centre of the eyes, the actual observer points of views (PoV), in the base reference frame (Σ_o) (see Fig. 3).

The Observer PoV

When the helmet is strapped to the head of the observer, both eyes occupy a fixed position in the helmet reference frame (Σ_t). In such a condition, the location of the centre of each eye can be mathematically described by a vector (${}^{t}P_{eye}$) relative to the helmet reference frame (Σ_t) (see Fig. 3). The coordinates of the eye in the helmet frame (${}^{t}P_{eye}$) can be further represented in the base reference frame (Σ_o) by referring to Eq. (1). Mathematically, the transformation can be written as:

$$ {}^{o}P_{eye} = {}^{o}_{t}T(\underline{q}) \cdot {}^{t}P_{eye} \qquad (2) $$

where ${}^{o}_{t}T(q)$ is the time varying matrix which transforms the coordinates of a point from the helmet reference frame (Σ_t) to the base reference frame (Σ_o). In the section The Position of the Eye Within the Helmet, we will describe a technique which can be used to estimate the position of the centre of the eye (${}^{t}P_{eye}$) in the helmet reference frame (Σ_t).

In order to generate, on a given projection surface, the virtual image corresponding to the actual PoV of the observer, it is necessary to express the eye position, the virtual scene and the projection surface in a common frame of reference. From a mathematical point of view, this implies the transformation of the coordinates of the eye position (${}^{o}P_{eye}$), as described by Eq. (2), into the reference frame of the projection surface (Σ_s). This additional transformation of coordinates, completes the so called *visual display* transformation [22] which maps, according to the momentary observer PoV, points of the virtual scene into the reference frame of the projection surface (Σ_s). We discuss in the next section how to formalise mathematically this transformation of coordinates.

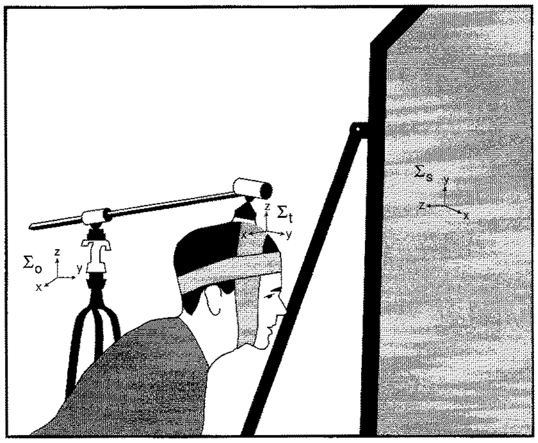

Fig. 4. Experimenting in large-field, immersive visual environment. The motion capture device provides the position information of the centre of the eye in the base reference frame (Σ_o). The eye coordinates need to be further transformed into the reference frame (Σ_s) integral to the large screen in order to generate correct images of a virtual scene.

Observer PoV in Screen Coordinates

In 3D viewing, an image of a virtual object is generated by specifying a centre of projection (i.e. the observer PoV) and a projection plane (i.e. its position and orientation) [23]. The projection of a 3D object is so forth defined by straight projection rays (called projectors) emanating from the centre of projection, passing through each point of the modelled object, and intersecting the projection surface. A common frame of reference is needed for the observer PoV, the projection surface and the virtual object. Therefore, we need to transform the coordinates of the observer's PoV, from the reference frame integral to the base (Σ_o), to the reference frame integral to the projection surface (Σ_s) (see Fig. 4).

The relationship between the base of the device and the projection surface is not known a priori and needs to be estimated. This relationship can be represented by a 4×4 matrix ${}^{s}_{o}T$. The matrix, once identified, will enable to generate in response to a motor behaviour of the observer, a coherent visual feedback. Mathematically, the 4×4 matrix combined with Eq. (2) gives:

$$ {}^{s}P_{eye} = {}^{s}_{o}T \cdot {}^{o}_{t}T(\underline{q}) \cdot {}^{t}P_{eye} \qquad (3) $$

which provides the coordinates of the eye of the observer (${}^{s}P_{eye}$) in the screen reference frame (Σ_t). In the next section, we show how to identify the coordinate transformation ${}^{s}_{o}T$, using the measuring abilities of the motion capture device.

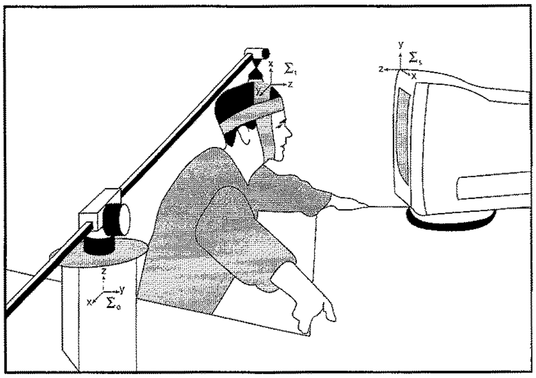

Fig. 5. Experimenting in small-field, non-immersive environment. In this configuration, the projection surface (Σ_s) is the desktop monitor. Once the position of the base is localized with respect to the screen, a procedure to identify the visual display transformation matrix ${}^{s}_{o}T$ is required.

Estimating the Visual Display Transformation

A direct method to estimate the transformation ${}^{s}_{o}T$ is to exploit the measuring capabilities of the motion capture device. The end part of the device was designed so as to easily attach and remove a pointer. The coordinates of the tip of the pointer (${}^{t}P_{pt}$) are known in the helmet reference frame Σ_t. Its linear dimensions have been calibrated during manufacturing. If we substitute in Eq. (2) the coordinates of the eye vector (${}^{t}P_{eye}$) with the pointer coordinates (${}^{t}P_{pt}$), the endpoint coordinates are also known in the base reference frame Σ_o, as results from the following:

$$ {}^{o}P_{pt} = {}^{o}_{t}T(\underline{q}) \cdot {}^{t}P_{pt} \qquad (4) $$

When configured with the pointer, the motion capture device becomes a 3D digitizing tool which enables us to sample the locations of reference points in the projection surface. Figure 6 shows the stylus pointer is mounted at the end point of the motion capture device.

The procedure used to estimate the matrix ${}^{s}_{o}T$ is based on the idea of matching a predefined set of reference points (P_i, i=1 . . . N) whose relative positions are known in the screen and the base reference frames, respectively, Σ_s and Σ_o. In order to perform correctly the matching, an accurate calibration fixture is used as shown in Fig. 6. The calibration fixture is composed of a rectangular frame containing a regular grid of thin metallic wires. The wires are equally spaced in the horizontal and vertical directions, forming a matrix of 7×9 squared tiles, measuring 27 cm on each side. The

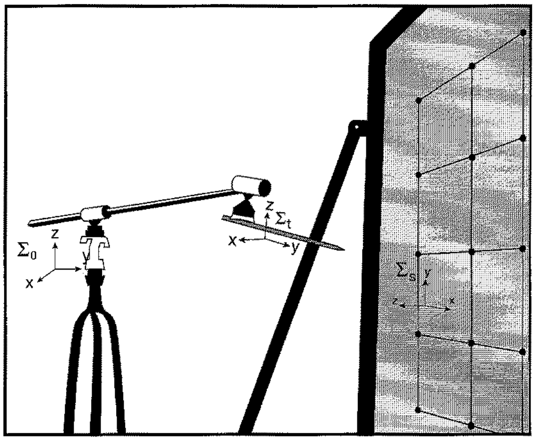

Fig. 6. The device transformed in a calibration tool. A stylus pointer can be fixed and removed to the end part of the device and transform it into a 3D calibration tool. In this configuration the device can be used to measure the 3D position of reference points with respect to the base reference frame Σ_o). A calibration fixture is superposed to the screen to define a set of reference points.

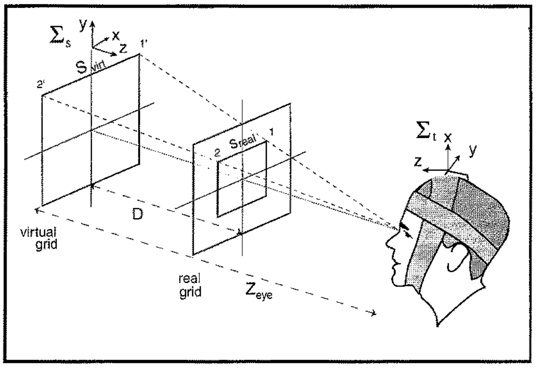

Fig. 7. Schematic diagram of the eye calibration procedure. Two grids are used. A real grid of known size (S_{real}) is positioned at a specified distance (D) from the screen surface. A virtual grid generated on the screen surface has a variable size (S_{virt}). The observer aligns his/her eye centrally with the two grids. The alignment constrains the values of two of the three Cartesian coordinates measured in the Σ_s frame. The third Cartesian coordinate (Z_{eye}) can be estimated using basic geometry between and the knowledge of the actual grids parameters (S_{real}, D, S_{virt}).

grid intersections define the set of reference points (P_i,i=1 ... N). The calibration fixture is superposed to the projection surface, so as to provide also a set of references for the alignment of the graphical output produced by the screen projector. The rotational and the translational parts of the transformation matrix ^s_oT are obtained from the coordinates of the reference point P_i measured in the Σ_o reference frame. In particular, the rotational part ^s_oR is derived by computing the *directional cosine matrix*, i.e. the matrix of the unit vectors denoting the direction X_s, Y_s, Z_s of the Σ_s frame expressed in the Σ_o reference frame. In the appendix, we describe one procedure to identify the elements of the directional cosine matrix. The translation vector (^s_oP) which describes the position of the origin of the screen reference frame (Σ_s) with respect to the base reference frame (Σ_o) is also derived from the P_i coordinates of the reference points.

A similar approach can be used to estimate the visual display transformation when the VE is configured with a desktop monitor (Fig. 5). In this latter case, the calibration fixture is not necessary, since a regular grid can be generated directly on the monitor surface using simple graphic primitives.

The Position of the Eye Within the Helmet

The position of the centre of the eye in the helmet ($^tP_{eye}$) (see Fig. 3) can be estimated using a simple alignment procedure which is performed by each individual subject. A schematic representation of the alignment procedure is outlined in Fig. 7. Two grids are used as optical alignment references. The first grid is real. It is composed of thin metallic wires and is accurately positioned in front of the screen. The second grid is a virtual image generated on the screen. The actual size (S_{virt}) of this latter grid can be linearly scaled using a keyboard manual control. As for the real grid, its linear dimensions (S_{real}) are known and is positioned at a specified distance (D) from the projection surface. Both grids are parallel and aligned in their central part.

The coordinates of the centre of the eye ($^tP_{eye}$) in the helmet frame (Σ_t), can be equivalently described in the screen reference frame (Σ_s), as shown by Eq. (3). When the subject aligns his eye with the centres of the two grids, the position of the eye measured in the screen frame ($^sP_{eye}$) is constrained to lay on the axis passing through the two grid centres. The eye coordinates will assume the form: (0,0,Z_{eye}). *The procedure continues as follows:*

1. the observer, using a chinrest, aligns one of the two eyes with the central part of the two grids. This alignment constrains the eye position measured in the screen frame of reference (Σ_s) to a family of Cartesian coordinates of the type: (0,0,Z_{ey});
2. the observer scales the actual size (S_{virt}) of the virtual grid with a manual keyboard control. Once the outer frame of the virtual grid disappears

behind the real grid, achieving the optical alignment, the actual scaled size (S_{virt}) is confirmed using a second manual keyboard control. The confirmed size (S_{virt}) is used to estimate the actual distance of the eye (Z_{eye}) from the screen surface. Under the hypothesis of perfect optical alignment, the distance variable obeys the following geometric relationship:

$$Z_{eye} = D \frac{S_{virt}}{(S_{virt} - s_{real})} \qquad (5)$$

The two-step procedure is repeated a predefined number of times (N), each time with the head oriented in a slightly different way. Each alignment-and-scaling trial provide one set of measurements for the following equation

$$^{s}P_{eye} = {}^{s}_{o}T \cdot {}^{o}_{t}\mathrm{T}(\underline{q}_k) \cdot {}^{t}P_{eye} \qquad (6)$$

where $^{t}P_{eye}=(x_{eye},y_{eye},z_{eye},1)$ is the unknown homogeneous variable and $^{s}P_{eye}=0,0,Z_{eye},1)$ is the leftmost term of Eq. (6). The repeated measurements are used to solve a system of linear equations using least-square methods. This system of equation, for which the unknown variable is $Y={}^{t}P_{eye}$, can be described as:

$$u=B \cdot [\mathrm{Y}]$$

where

$$u = \begin{bmatrix} ^{s}P_{eye}{}^{\cdot}1 \\ ^{s}P_{eye}{}^{\cdot}N \end{bmatrix} \qquad B = \begin{bmatrix} {}^{s}_{o}T \cdot {}^{o}_{t}T(\underline{q}_1) \\ {}^{s}_{o}T \cdot {}^{o}_{t}T(\underline{q}_N) \end{bmatrix}$$

are respectively a vector of dimension $3N \times 1$ and a matrix of dimension $3N \times 4$. The term $^{s}P_{eye_i}$ (I=1 ... N) represents the i-th measurement of the N predefined set of measurements.

Experiments in Large- and Small-Field VEs: Two Example Studies

This section will describe how the VE framework has been successfully used to design two experimental investigations addressing the study of 3D visual perception of a moving an observer as compared with the one experienced by a stationary observer. The first study, was conducted in large-field, fully immersive conditions. Its aim was to show that the central nervous system (CNS) uses self-motion-related signals to estimate the egocentric distance of an object from monocular, optic flow cues [24,25]. The second study was conducted in small-field conditions. In this case, the goal was to show that the CNS makes use of self-motion-related signals to interpret the visual image when conflicting visual cues are simultaneously displayed [26]. Central to both experimental protocols is the possibility to recreate for the stationary observer an accurate replay of the visual stimulation experienced by the active observer. The characteristics of the VE framework such as absence of latency and high accuracy are fundamental to the achievement of such a central issue.

Large Field VE

The rationale for the first investigation is briefly reviewed. An observer who moves in a real environment experiences on his/her retina many depth cues. Among these cues the one which is considered among the most effective is *motion parallax* [27]. In order to take advantage of this potent depth cue, it is not necessary to move the whole body, but a simple head movement is sufficient [28]. From theoretical work [29,30] it is known that a pattern of movement on the retina, i.e. a retinal optic flow, is informative of the 3D structure of the environment only up to a scaling factor [29,31]. Computationally speaking, the analysis of an optic flow by a stationary observer cannot lead to estimation of absolute distances of objects. Conversely, an active observer is in principle able to recover the absolute distance of an object from the pure optic flow, if his/her CNS exploits the measurement of the 3D movement accompanying the retinal motion experienced.

The first study tested the hypothesis that the CNS makes use of self-motion-related signals to scale the optic flow for determining the egocentric distance of objects [24]. The VE framework was integrated to a large-field visual display comprising a high-resolution projector (BARCO) displaying images of 1280×1024 pixels on a 200×250 cm screen. Two experimental conditions were generated, which enabled subjects to experience the same visual cue, i.e. motion parallax, under two different motor conditions. In the first condition, motion parallax was actively generated by the observer who moved the head side to side (the *self-motion* condition). In the second condition, the same motion cues were generated by the movement of the objects, which were passively experienced by the

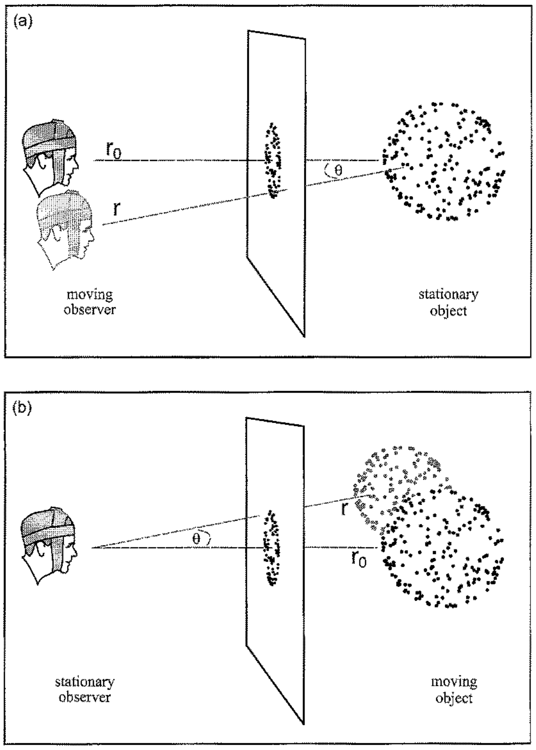

Fig. 8a,b. Schematic representation for the study of 3D layout. (a) The active condition in which the subject voluntarily controls his movement. (b) The passive condition in which subject observes from a stationary position. In the two conditions, the relative displacement between the observer and the object ($r-r_0$) is identical, therefore subjects experience the same optical stimulation.

stationary observer (the *object-motion* condition). Objects were generated at different distances (Fig. 8a), and relative size cues were removed by co-varying the object simulated size and the object actual distance. In the second condition, the stationary observer saw an object which moved along a trajectory identical to the one performed by the head of the active observer (Fig. 8b). The relative movement between the observer and the object was the same in both conditions, therefore from a visual point of view, the same optical stimulation (i.e. the same optic flow) was experienced. In the second condition, though, none of the information about self-motion (i.e. proprioceptive, vestibular, efference copy of motor command) was available.

The method used to generate in the passive condition, an optical stimulation identical to the active condition, is based on the equality of relative displacements for equal time intervals. During each trial (active condition), the instantaneous trajectory $r(t)$ of the head was recorded starting from the initial position r_0 (see Fig. 8a). During the corresponding passive trial, the object was moved along a trajectory which is the replay of the relative displacement recorded in the active condition. If we indicate by r_0 the eye position at time t_0 and by r the eye position at a given time (t) during the subject movement, then in the passive condition, at the corresponding time (t) the object centre (O) was displaced from its initial position (O_0) (see Fig. 8b) by an amount resulting from the following:

$$O=O_0-(r-r_0) \tag{7}$$

Therefore, because in both conditions the relative movement between the observer's eye and the virtual object was identical, if the observer tracked the object (he/she was instructed to do it), in the passive condition at time instant t the optic flow produced on the observer's retina was identical to the motion parallax experienced during the active condition.

The results of two studies [24,25] have demonstrated that the CNS exploits extra-retinal signals elicited by self-motion to provide an absolute scale to the motion parallax visual information. Subjects were able to recover correctly the absolute distance of objects while moving their head (Fig. 8a). Conversely, they failed to recover the absolute distance to the object, when they observed from a stationary position and could not weight the optic flow cues against self-motion signals (Fig. 8b).

Small-Field VE

The second investigation was aimed at testing how the CNS weights two different visual cues to depth, respectively linear perspective and motion parallax, in two alternative observation conditions: stationary observation and while moving. The visual task consisted of specifying the 3D orientation of a regular grid seen in the two observation conditions. In the first condition, the observer looked at a stationary grid while performing side-to-side head movements (Fig. 9a). The observer's head movement produced the visual motion on the retina. In the second condition, the observer watched from a stationary position the planar grid in motion (Fig. 9b). In this case, the object movement produced the visual motion on the retina. In both conditions, conflicts between the two visual cues were introduced so as to perturb the estimation of 3D structure from motion (see [26] for further details). In a given number of trials, the orientation of the grid surface defined by perspective cues differed by predetermined amounts with respect to the orientation defined by motion parallax cues.

As for the method, a procedure similar to the one described in the previous study was used for this

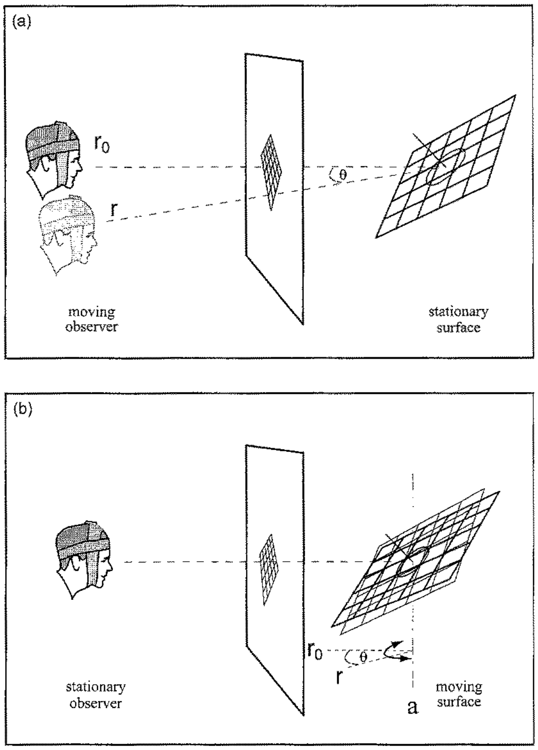

Fig 9a,b. Schematic representation of the study on perception of 3D orientation. (a) The active condition. (b) The passive condition. The knowledge of the momentary eye position of the observer with respect to the virtual object enabled us to reproduce, with high fidelity in the passive condition, the optic flow experienced by the observer during his own movement in the previous phase.

experiment. If we indicate by r_0 the position of the eye at time t_0 and by r the position of the eye at a given moment (t), during the subject movement, θ, the angle between r_0 and r is given by the following:

$$\theta = arcos[(r_0 \cdot r)/(\|r_0\| \cdot \|r\|)] \qquad (8)$$

and a, the instantaneous axis of rotation centred on the virtual object is defined as:

$$a = r_0 \times r \qquad (9)$$

In the passive condition, the observer remained stationary, while watching a replay of the same visual motion. This time the visual motion was produced by a rotation of the virtual object of an angle θ around an axis aligned with the vector a, passing through the centre of mass of the object. The rotation is such that the optic flow seen by the stationary observer was identical to the one experienced during a preceding active trial.

Moreover, in both cases when the observer was stationary, he was still wearing the helmet; therefore any small involuntary movement could be taken into account to compensate the projection of the virtual object. Contrary to the preceding case, here the object was simply rotated and not translated.

Regarding the first study, the results suggest that self-motion information is actually incorporated into visual judgments of three-dimensional structure and that in general it is inadequate to exclude observer motion in the study of 3D visual perception.

Conclusion

VE technology is becoming an extremely interesting tool for fundamental and applied research. In the neuroscience field, for example, the technology has been already exploited in several investigations of sensory-motor and cognitive functions of the human brain [5,6,8,9,32]. In applied research, on the other hand, VR technology is becoming very interesting for engineering design, prototypical manufacturing, technological and human factors studies. For example, VE interfaces are nowadays used during the design process of new products and as an effective tool in training [33]. In the automotive industry, VR techniques are exploited to evaluate different technological solutions, for example to develop new intelligent lighting systems [34].

In this work, we have described a VE framework, which was conceived to design experiments in visual perception and action. Its main characteristics stem from the design of a six degrees of freedom motion capture device [21]. The operation of this device is based on mathematical modelling of its multi-link structure according to methodologies used in robotics [20]. As a consequence its operation is largely insensitive to interfering field and metal, which is not the case for electromagnetic trackers [35]. The device is highly accurate in position measurement and has almost no latency in the tracking position, is flexible with respect to the design of monocular or binocular experiments and it is adaptable to different display solutions (large and small field). Last, but not least, it is easy to use.

Using the VE framework described in the paper, two action-perception studies have been accomplished. These studies compared the performance of an active and a passive observer in perception of 3D space. The studies have demonstrated that perception of 3D space is much more robust when multi-modal sensory information is integrated, with particular reference to visual and self-motion information. For example, absolute distances are more reliably estimated by a moving observer [24,25] than by a passive observer (i.e. an observer who stands still and experiences the same optical stimulation). Furthermore, some of the spatial attributes of objects in the environment (e.g. stationarity) are better judged by a moving observer than by a passive one [26].

These kinds of study provide insights into are the relevant sensory information that the brain relies on for 3D space perception. In particular, they show that action within a VE is a very important trigger to robust perception of 3D space. These insights should be helpful to designers of VEs, suggesting that action is to be considered an important factor in design of VR systems. For example, the user performance in tasks that require some metric evaluation within the simulated environment (e.g. grasping or evaluating the distance of objects) could be potentially improved if the VR technology enables head movement in natural conditions. Furthermore, natural movement within a VE has been reported to be one important factor favouring immersion [36]. In the first of the two studies, 10 subjects out of 12 verbally reported a high level of immersion when performing visual exploration under large-field conditions with natural head movements. Conversely, the same subjects reported a moderate level of immersion when watching from a stationary position the same optical stimulation. Hopefully, the type of VE framework we have described, and the studies that are made possible by its exploitation, might contribute to better focus the efforts related to the study of human performance efficiency in VR systems.

Acknowledgement

The authors gratefully acknowledge France Maloumian for her help in preparing the figures.

References

1. Riva G (1998) Virtual environments in neuroscience. IEEE Transactions on Information Technology in Biomedicine 2: 275–281
2. Pelz JB, Hayhoe MM, Ballard DH, Shrivastava A, Bayliss J, and von der Heyde M (1999) Development of a virtual laboratory for the study of complex human behavior. In: The engineering reality of virtual reality. Proceedings of the SPIE, San Jose, CA 3639B
3. Findlay JM and Newell FN (1994) Perceptual cues and object recognition. In: Simulated and virtual realities: elements of perception. Carr K and England R eds. London: Taylor & Francis 113–130
4. Merril JR (1997) Using emerging technologies such as virtual reality and the World Wide Web to contribute to a richer understanding of the brain. Annals of the New York Academy of Sciences 820: 229–233
5. Ghahramani Z, Wolpert DM (1997) Modular decomposition in visuomotor learning. Nature 386: 392–395
6. Aguirre GK, Detre JA, Alsop DC, D'Esposito M (1996) The parahippocampus subserves topographical learning in man. Cerebral Cortex 6: 823–829
7. Bertin RJ, Israel I, Lappe M (2000) Perception of two-dimensional, simulated ego-motion trajectories from optic flow. Vision Research 40 (21): 2951–2971
8. Viaud D, Ivanenko YP, Berthoz A, Jouvent R (1998) Sex, lies and virtual reality. Nature Neuroscience 1: 15–16
9. Amorim MA, Trumbore B and Chogyen PL (2000) Cognitive repositioning inside a 'desktop' VE: the constraints introduced by first- vs. third-person imagery and mental representation richness. Presence 9: 165–186
10. Kennedy R, Lane N, Lilienthal M, Berbaum K, Hettinger L (1992) Profile analysis of simulator sickness symptoms: application to virtual environment systems. Presence 1: 295–301
11. McCauley ME, Sharkey TJ (1992) Cybersickness: perception of self-motion in virtual environment. Presence 1: 311–318
12. Stanney KM, Hash P (1998) Locus of user-initiated control in virtual environment: influences on cybersickness. Presence 7:447–459
13. Land MF (1992) Predictable eye-head coordination during driving. Nature 359: 318–320
14. Grasso R, Glasauer S, Takei Y, Berthoz A (1996) The predictive brain: anticipatory control of head direction for the steering of locomotion. Neuroreport 7: 1170–1174
15. Koch C, Poggio T (1999) Predicting the visual world: silence is golden. Nature Neuroscience 2: 9–10
16. Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2: 79–87
17. Harris LR, Jenkin M, Zikovitz DC (2000) Vestibular capture of the perceived distance of passive linear self motion. Archives Italiennes de Biologie 138: 63–72
18. Harris LR, Jenkin M, Zikovitz DC (1998) Vestibular cues and virtual environments. In: Proceedings of VRAIS 1988. Atlanta, GA: IEEE 133–188
19. Hendrix C, Barfield W (1995) Presence within virtual environments as a function of visual display parameters. Presence 5: 263–273
20. Yoshikawa T (1990) Foundations of robotics: analysis and control. Cambridge, MA: MIT Press
21. Panerai F, Hanneton S, Droulez J, Cornilleau-Pérès V (1999) A 6-dof device to measure head movements in active vision experiments: geometric modeling and metric accuracy. Journal of Neuroscience Methods 90: 97–106
22. Robinett W, Holloway R (1995) The visual display transformation for virtual reality. Presence 4: 1–23
23. Foley JD, van Dam A, Feiner SK, Hughes JF (1996) Computer graphics: principles and practice. Reading, MA: Addison-Wesley
24. Panerai F, Cornilleau-Pérès V, Droulez J (2002) Contribution of extra-retinal signals to the scaling of object distance during self-motion. Perception & Psychophysics
25. Peh CH, Panerai F, Droulez J, Cornilleau-Peres V, Cheong LF (2002) Absolute distance perception during in-depth head movement: calibrating optic flow with extra-retinal information. Vision Research (in press)
26. Wexler M, Panerai F, Lamouret I, Droulez J (2001) Self-motion and the perception of stationary objects. Nature 409: 85–88
27. Sekuler R, Blake R (1994) Perception. New York: McGraw-Hill

28. Rogers B, Graham M (1979) Motion parallax as an independent cue for depth perception. Perception 8: 125–134
29. Lee DN (1980) The optic flow field. Philosophical Transactions of the Royal Society of London (B) 290: 169–179
30. Koenderink JJ (1986) Optic flow. Vision Research 26: 161–180
31. Prazdny K (1983) Information in optic flows. Computer Vision, Graphics, and Image Processing 22: 235–259
32. Goodbody SJ, Wolpert DM (1999) The effects of visuomotor displacement on arm movement paths. Experimental Brain Research 127: 213–223
33. Flipo A (2000) TRUST: The truck simulator for training. In: Proceedings of Driving Simulation Conference (DSC2000). Paris: INRETS 293–302
34. Dubrovin A, Levelé J, Prevost A, Lecocq P, Canry M, Kelada JM, Kemeny A (2000) Application of real-time lighting simulation for intelligent front-lighting studies. In: Proceedings of Driving Simulation Conference (DSC2000). Paris: INRETS 333–343.
35. Nixon MA, McCallum BC, Fright WR, Price NB (1998) The effects of metals and interfering fields on electromagnetic trackers. Presence 7: 204–218
36. Witmer BG, Singer MJ (1998) Measuring presence in virtual environments: a Presence questionnaire. Presence 7: 225–240

Correspondence and offprint requests to: *Francesco Panerai, Laboratoire de Physiologie de la Perception et de l'Action (LPPA), CNRS/Collège de France, 11 pl. Marcelin Berthelot, F-75005 Paris, France. Email: francesco.panerai@college-de-france.fr*

Appendix

Homogeneous Coordinates to Describe Multi-Link Structures

In order to represent mathematically the position and the orientation of the end part of the mechanical device, the concepts of coordinate frame and coordinate transformation (i.e. transformation between adjacent frames) are required. A convenient choice to describe coordinate transformations between the adjacent frames (i.e. the adjacent links in a multi-link structure) is using *homogeneous coordinates*. For example, the relationship between two generic frames, the frame Σ_1 and the frame Σ_2 shown in Fig. 10, can be described by a combination of a translation (i.e. vector) and a rotation (i.e. matrix). The rotational matrix and the translational vector can be assembled in a 4×4 matrix defining completely the relation between Σ_1 and Σ_2:

$$ {}^1_2T = \begin{pmatrix} {}^1_2R & {}^1P_2 \\ 0 & 1 \end{pmatrix} $$

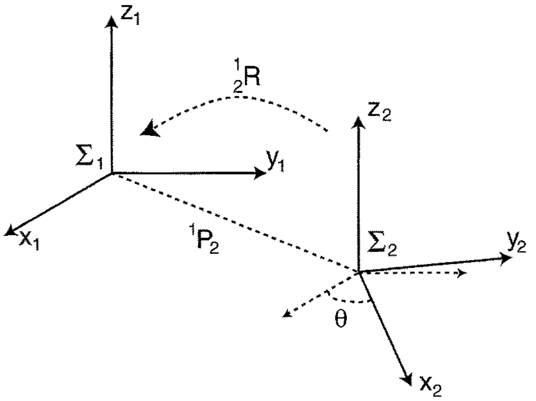

Fig. 10. Coordinate transformation between two generic frames. The position of the frame Σ_2 with respect. to Σ_1 is specified completely by the position vector ${}^1P_2$ and the rotation matrix ${}^1_2R$.

where ${}^1_2R$ is the 3×3 matrix defining the rotation, θ, of frame Σ_2 with respect to Σ_1, and ${}^1P_2$ is the vector defining the position of the origin of frame Σ_2 with respect to Σ_1.

Description of complex multi-link structures can be accomplished by composing several matrices such as ${}^1_2T$ [20].

The Visual Display Transformation

We describe the procedure used to identify the transformation matrix ${}^s_oT$. The procedure is based on the measurement of a set of reference points $P_{ij}(i=1 \ldots n, j=1 \ldots m)$ integral to the projection surface. The points are defined by the intersection of a 9×7 grid of thin metallic wires, which is superposed to the projection surface by means of a calibration fixture. The transformation matrix ${}^s_oT$ enables generation of the correct viewing projection corresponding to the momentary point of view (PoV) of the observer. It transforms the coordinates of the observer PoV from the frame in which the movement is measured (Σ_o) to the frame in which the image is generated (Σ_s). According to the previous section, the transformation can be represented as:

$$ {}^1_2T = \begin{pmatrix} {}^1_2R & {}^1P_2 \\ 0 & 1 \end{pmatrix} $$

where ${}^s_oR$ is the rotation matrix from Σ_s to Σ_o defined as:

$$ {}^s_oR = [{}^o\underline{x}_s \;\; {}^o\underline{y}_s \;\; {}^o\underline{z}_s]^T $$

and ${}^{s}_{o}O$ is the vector representing the position of the origin of Σ_o with respect to Σ_s and results from:

$${}^{s}O_0 = -\frac{1}{NM}\sum_{ij} P_{ij}$$

where N=7 and M=9. The unit vectors of the viewing frame (Σ_s), i.e. (${}^{o}x_s$ ${}^{o}y_s$ ${}^{o}z_s$, are defined as follows: (1) the z-unit vector is the vector (n) orthogonal to the projection surface, which is defined as the plane interpolating all the reference points $P_{ij}(i=1 \ldots 7, j=1 \ldots 9)$; (2) the y-unit vector is the normalised vector $\underline{\upsilon}$, obtained from the average coefficient of the 3D vertical meridians of the calibration fixture. The x-unit vector is obtained by applying the Gram–Schmidt orthonormalisation technique to the previous unit vectors. In mathematical terms, the unit vectors in the directions X_s, Y_s, Z_s expressed in the Σ_o frame can be written as:

$$\underline{n} = |n_x, n_y, n_z| = \frac{(A \cdot A^T)^{-1} \cdot \underline{b}}{\|(A \cdot A^T)^{-1} \cdot \underline{b}\|} \qquad \underline{b} = [1..1]^T, A = [P_{ij}^T]$$

and

$$\underline{\nu} = |l, m, n| = \frac{(C \cdot C^T)^{-1} \cdot \underline{d}}{\|(C \cdot C^T)^{-1} \cdot \underline{d}\|} \qquad \underline{d} = [(-\underline{n}^T \cdot \underline{X}_{ij}), C = [\underline{t}_j \cdot \underline{n}^T]$$

$t_j = \frac{j}{N}$ being the coefficient identifying the j-th reference points along a given vertical meridian.