SAM
Science Arts & Métiers
Archive ouverte Arts & Métiers ParisTech
ARTS ET MÉTIERS ParisTech

To cite this version :

Mohammad Ali MIRZAEI, Jean-Rémy CHARDONNET, Frédéric MERIENNE, Christian PERE - Improvement of the real-time gesture analysis by a new mother wavelet and the application for the navigation inside a scale-one 3D system - In: 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Poland, 2013-08-27 - 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance - 2013

Improvement of the real-time gesture analysis by a new mother wavelet and the application for the navigation inside a scale-one 3D system

M. Ali Mirzaei, Jean-Rémy Chardonnet, Frédéric Mérienne, Christian Père
Arts et Métiers ParisTech, CNRS, Le2i, Institut Image, Chalon-sur-Saône, France
mirzai142.nri@gmail.com

Abstract

This paper proposes a navigation technique for traveling inside a real-scale 3D model based on human gesture analysis. In the first step, a simple threshold is used as a criterion to analyze gestures. In the next step, a complex criterion will be imposed to the analysis to improve the navigation technique. Walking is a periodic signal and moving feet up and down is a part which is repeated. A mother wavelet is allocated to the selected pattern. Then the position of the pattern is recognized by applying Multi Resolution Analysis (MRAs). The movement command is generated and sent to the graphic render (as a movement command). Analytical results show a very high precision performance in the presence of noise, scale variation and superposition of other signals. Practical experiments also verify the same promising results.

1. Introduction

Designing efficient interaction and navigation for traveling inside a real-scale (scale one) 3D model, so called Immersive Virtual Environment (IVE), has been a research topic for nearly three decades. Recent innovations in the field of digital video technology, new generation of infrared (IR) sensors and measurement instruments have changed the orientation of human machine interface (HMI) researches from simple mechanical [7] to wireless [15] and touch-less [6] devices. Sensors play an important role in developing this modern end-user HMI in IVEs, especially when the aim is to develop an interface (interaction/navigation) for a real-time system.

One long-term attempt in HMI has been to develop and integrate the “natural” means that humans employ to communicate with each other into HMI. Human gesture provides a way to interact with and navigate in real-scale 3D systems such as virtual reality systems [8] and IVEs [5]. Head detection using markers [16], hand tracking [2], and locomotion perception [13] are some of the steps that have been taken to facilitate this matter. Among different vision sensors, infrared has a great deal of importance due to its precision in depth estimation.

Infrared cameras in a multiple view configuration can send out very high precision details about the depth and the location of the object by means of traveling wave techniques but it is an expensive solution. Besides, the calibration and the installation of the system is time consuming. On the contrary, a single IR camera-projector configuration (such as Kinect) is simple to calibrate and extremely cheap [11, 12]. However, it can acquire data only for ordinary applications which require fairly low resolution and precision. For high precision applications, either higher resolution RGB camera needs to be combined with an IR camera-projector configuration, or post processing should be applied to the output signal.

Human skeleton detection [9] by using different joints or pair joints was one of the common way to use human gestures in HMI design. For example, consecutive up/down movement of the right and left ankles can be acquired by IR camera and interpreted as a real-time walking signal. Then some features of the signal (max, min, etc.) can be extracted by signal processing approaches to command a 3D system to move forward/backward and turn to the left/right [20]. The position of the hand and the configuration of the fingers can be used to interact with a 3D object in an IVE.

Our contribution is to propose and assess a navigation method for walking inside a scale-one 3D system based on gait analysis by feature matching and adapting mother wavelet from a walking pattern. The scale-one 3D system refers to a virtual reality HMI in which people walk and their walking gait is used for navigation in an IVE.

The paper is organized as follows: the principle of depth measurement and the calculation of the object coordinates will be explained in section 2. Walking signal extraction and gait interpretation will be introduced in section 3. Hardware and software requirements of the test bench for the development and verification will be described in section 4. In section 5, the principle of multi-resolution analysis will be summarized. The last section is dedicated to experimental results and the performance analysis.

2. Depth signal and object coordinates calculation using infrared camera

The relation between the distance of an object k to the sensor relative to a reference plane and the measured disparity d are shown in Figure 1. A depth coordinate system with its origin at the focal point of the infrared camera will be established to express the 3D coordinates of the object points. The Z-axis is orthogonal to the image plane towards the object, the X-axis perpendicular to the Z-axis, and the Y-axis orthogonal to X and Z making a right handed coordinate system. Assume that an object is located in the reference plane at a distance Z_0 and a point on the object is projected on the image plane of the infrared camera. If the object is shifted closer to the camera, the new location of the point on the image plane will be displaced along the X-direction (disparity d). Substituting D from $D/b = (Z_0 - Z_k)/Z_0$ (from the similarity of ΔOCL and ΔOMK triangles) into $d/f = D/Z_k$ (from the similarity of ΔHKC and ΔCGA triangles) and expressing Z_k in terms of other variables yields $Z_k = Z_0/\left(1 + \frac{Z_0}{fb}d\right)$, where Z_k denotes the distance (depth) of point k in the object space, b is the base length between the camera and the projector, f is the focal length of the infrared camera, D is the displacement of point k in object space, and d is the observed disparity on the image plane. The constant parameters Z_0, f, and b can be determined by the camera calibration. The Z coordinate of a point together with f defines the imaging scale for that point. The coordinates of each point on the object can then be calculated from its image coordinates and the scale [19]:

$$X_k = \frac{Z_k}{f}(x_k - x_0 + \delta x) \tag{1}$$

$$Y_k = \frac{Z_k}{f}(y_k - y_0 + \delta y) \tag{2}$$

Where x_k and y_k are the image coordinates of the point, x_0 and y_0 are the coordinates of the principal point, and δx and δy are corrections for lens distortion.

3. Gesture analysis

3.1. Walking signal definition

The movement of a user is limited to a small area in a multi-projector real-scale 3D system. Therefore, the user is not free to walk more than few meters. Throughout the literature when we are talking about walking, it means the user will stay in a specific coordinate and will walk in place. When the user is walking in place, the left and right ankles are moving up and down one after another. If only the motion of an attached point to the right ankle is captured by an IR-camera and the X, Y, and Z coordinates of the point calculated by Eqs. (1) and (2), the variation signal along the Y axis (the Y-axis is in accordance with Figure 1) can be depicted as in Figure 2.b. The signal is called "walking with a constant step length". The difference between two picks shows the step length in Figure 2.b. The user might move few centimeters along the Z-axis (forward/backward) due to involuntary movement of the body. Movement in the Z-direction changes the amplitude of the signal. When

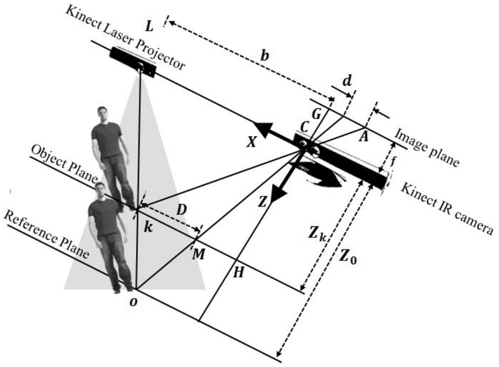

Figure 1. Schematic representation of the depth-disparity relation.

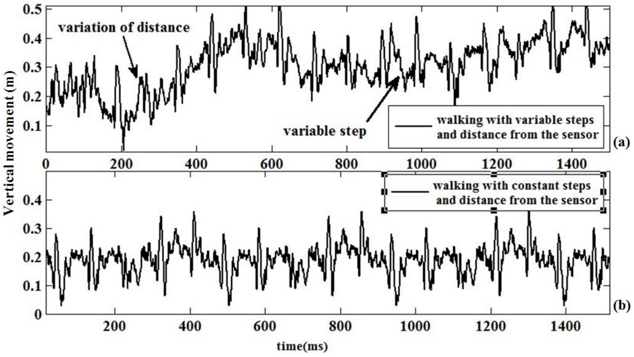

Figure 2. Example of a human walking signal with a) variable, b) constant steps and distance from the sensor.

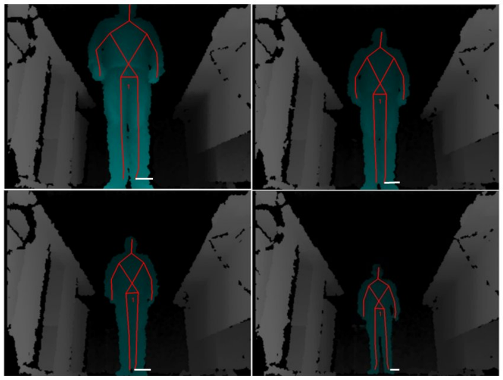

Figure 3. Scale variation (movement in the Z-direction) during walking in a real-scale 3D model.

the user is moving involuntarily along the Z-axis, the scale of the body in the images changes and it makes the same movement of the ankle look smaller or bigger. The distance between ankles and the ground is highlighted for four scales on Figure 3 by a small white line near the user's ankle. That is the reason why the amplitude of the signal is changing by the movements along the Z-axis (indicated in Figure 2.a). This can be deducted from Eq. (2) as well.

As involuntary movement of the body is controlled by an autonomous part of the Central Nervous System (CNS) [1], it is very hard to control this movement, especially when the user focuses on a navigation task. The movement along the X-axis has the same effect. Unlike Figure 2.b, the steps have different lengths in Figure 2.a. Moreover, when the ankle comes up more, the amplitude of the signal will be bigger and the length of the step grows. For instance, in Figure 2.a, around sample 1000, a longer step has been taken.

3.2. Navigation command generation based on walking signal analysis

The navigation task is constructed either by forward/backward movements and turning to the left/right, or by the combination of these movements and the itinerary. It means, to go from point A to C by the way of point B, we go from point A to B while using only translational (forward/backward) and rotational (left/right) movements, and then continue from B to C in the same way. Thus, the very basic navigation task is to go from point A to B by combining translational and rotational movements. Different methods are employed to implement a navigation task. In general, it can be classified under two generic methods: 1) with navigation devices (fly-stick, game-pad, Wii mote, etc.), 2) based on user's gestures.

In the first approach, a button is allocated to each subtask (translation and rotation) and the task is implemented by a function. By pushing each button the associated function is activated and the task begins. The function will run until the termination condition is met either by pressing another button or by checking a variable status.

In the second approach, the user's gesture is interpreted and the result is coded into a value. Then, the different values of the variable are used to initiate or terminate different tasks. A simple interpretation of the gesture will be introduced here just to make the methodology clear. More complicated analysis will be presented later (section 5). The current interpretation uses the threshold as an evaluation criterion. The analysis applies the criteria to the walking signal to generate a command pulse. The user's walking signals are shown in black line in Figure 4. The dot-line presents an activation threshold (depicted as "Walking signal", at the top of Figure 4). The forward function will be activated if the position of the ankle reaches the threshold level and will remain active till the ankle comes below this threshold. During a short period of time, a pulse will be activated. The length of the pulse is equal to the step length. The pulse activates and deactivates forward/backward movements. The acceleration of the scene movement is adjusted by the average value of the acceleration signal during the time pulse is active. In this interpretation, if the user takes steps with longer lengths, the scene will move more because the pulse active time is longer. The forward/backward signal comes from the orientation of the head. Rotation to the left/right is activated with the hand movement.

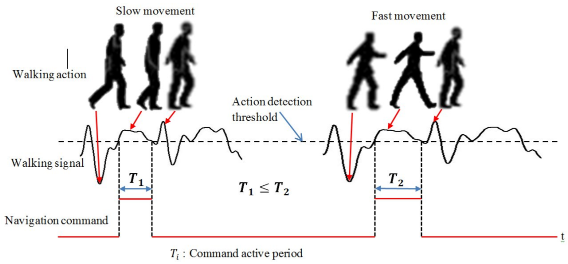

Figure 4. Sequence of navigation using gait interpretation.

4. Implementation and test apparatus

A CAVE^{TM} [4] system is used to implement and test the navigation system. This system consists in four walls, two projectors per wall (one image per eye, totally two images per face). An infrared based head tracking system (AR-tracker [21]) is used to find the user location. A custom platform called PeTRIV was developed to manage the connection between display projectors, infrared cameras and the networking. The platform uses OpenSceneGraph on the top of OpenGL to render the 3D model. Then the model is projected into the display system by MPI and four NVidia Quadroplex GPUs.

VRPN (Virtual-Reality Peripheral Network), proposed and implemented by Russell M. Taylor [18], has been used to connect an infrared camera-project setup (a Kinect Xbox 360 was the configuration for image grabbing and depth perception) to the display system and the graphic engine. A VRPN server developed by [17], widely known as FAAST, was employed to extract the skeleton and record the walking signal. This VRPN server provides the coordinates and the orientation of 24 joints of the user's body.

5. More precise gesture analysis by wavelet MRAs and pattern matching

It is very difficult to interpret user gestures by a simple threshold value. The amplitude of the signal will change when the user moves along the Z-direction as mentioned above. Therefore, if a high threshold is selected, some parts of the signal will remain under the threshold and the forward function will never be initiated. To avoid this problem, another approach will be selected to adjust the time and the activation period of the function. The activation pulse can be produced by defining a walking pattern and finding the pattern in the signal by a signal processing approach. Walking can be approximately explained by the pattern shown in Figure 5. Two walking patterns with different scales have been shifted and superimposed on a triangular wave and then they were recognized by wavelet analysis. The primary result shows that this method can find the position of the pattern very precisely. To show how wavelet decomposition is working, we will review the basics of wavelet analysis below.

In a wavelet multi-resolution analysis (MRAs) [3], a signal $x(t) \in V_1, V_1 \subseteq L(\mathbb{R})$ can be decomposed into a linear combination of an infinite series of detail functions, $\{G_1, G_2, ..., G_n\}$, so that

$$x\left(t\right) = \sum_{k=0}^{\infty} a_k G_k\left(t\right) \tag{3}$$

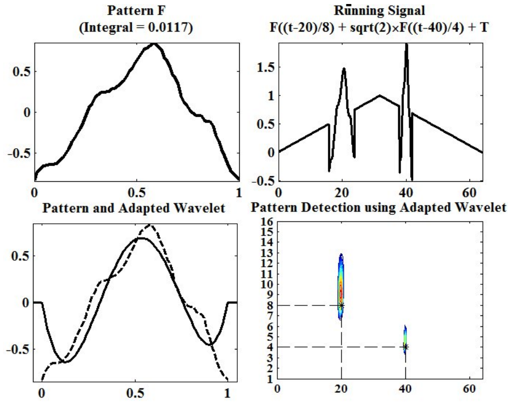

Figure 5. Generating a mother wavelet from the walking pattern.

Where a_k is expansion coefficients. Using the same method and just using different functions, each signal can be decomposed into detail and the approximation. To do this operation practically, the signals are projected on the base of space G_k. Different sets have been proposed for the base of the space, however the following sets which are called mother wavelet and scale functions are more popular.

$$\left\{\psi_{j,k} = 2^{-\frac{j}{2}}\psi(2^{-j}t - k), k \in \mathbb{Z}\right\} \tag{4}$$

$$\left\{\phi_{j,k} = 2^{-\frac{j}{2}}\phi(2^{-j}t - k), k \in \mathbb{Z}\right\} \tag{5}$$

Then using Eq. (4) to Eq. (5) the projection is done by,

$$x_j\left(t\right) = \sum_{k=-\infty}^{\infty} d_j^k 2^{-\left(\frac{j}{2}\right)}\psi\left(2^{-j}t - k\right) \tag{6}$$

$$y_j\left(t\right) = \sum_{k=-\infty}^{\infty} c_j^k 2^{-\left(\frac{j}{2}\right)}\phi\left(2^{-j}t - k\right) \tag{7}$$

$$d_j^k = \left\langle x_{j-1}\left(t\right), \psi_{j,k}\right\rangle, c_j^k = \left\langle x_{j-1}\left(t\right), \phi_{j,k}\right\rangle \tag{8}$$

Where, d_k^j and c_k^j are the projection coefficients and $\langle ., . \rangle$ is the inner product in L^2. Wavelet functions are not necessarily limited to those that have already been proposed for MRAs.

For specific signal analysis it is better not to use general methods because the result will have better precision. Rather, the better way is to find and adapt the wavelet which fits to the application, for instance gesture analysis in this study.

Here we proposed a mother wavelet function to take specific features from a walking signal. This paper will use a technique called adapting wavelet from pattern which is very well explained and documented in [3].

In practice and for numerical calculations we generally know only one sampling of a function f over an $[a, b]$ interval. We have a finite set of values $\{(t_k, y_k)\}_{k=1,...,k}$ such that: a_k and $y_k \approx f\left(t_k\right)$. We consider a family $F = \{\rho_i\}_{i=1}^{N}$ linearly independent in $L^2(a, b)$, and we denote by V, the span vector space of F. Formulated for this finite set of pairs, the problem is seeking $\alpha = \{\alpha_i\}_{i=1}^{N}$ in $\mathbb{R}^N$ and thus $\psi = \sum_{i=1}^{N} \alpha_i \rho_i$ such that:

$$\sum_{k=1}^{N} \left[\psi\left(t_k\right) - y_k\right]^2 = \min_{\beta \in \mathbb{R}^N} \left\{ \sum_{k=1}^{N} \left[V_\beta\left(t_k\right) - y_k\right]^2 \right\} \tag{9}$$

such that

$$\int_{b}^{a} V_\beta dt = 0 \tag{10}$$

Where for β in $\mathbb{R}^N$, $V_\beta = \sum_{i=1}^{N} \beta_i \rho_i$. In other terms, this formulation says, a polynom can be fitted to a pattern such that it satisfies the general condition of a mother wavelet [3] while maximizing the output of the MRAs.

The degree of the polynom is limited to six in this application to reduce the load of calculation and adapt the method for real-time processing. Usually, polynoms with a lower degree create huge amount of errors. The degree of the polynom needs to be more than three for better performance and less error. The selected pattern and adapted mother wavelet are shown in Figure 5.

After the selection of a pattern, we try to fit a polynomial function to the pattern such that it satisfies the definition of mother wavelet function [14]. The adapted mother wavelet is used to decompose the signal with MRA to find the position of the pattern. Two diagrams in the right side of Figure 5 show the detection accuracy with the variation of the scale factor. As seen, the superposition of the pattern with scale factors of four and eight has been successfully detected.

6. Experimental results and discussion

Walking and the coefficient signal of the adapted mother wavelet are shown in Figure 6.a,b. The local maximum and minimum are detected by MRAs analysis. MRA can be considered as a convolution of the pattern and the walking signal. When it is maximum, highly the pattern is located in that point. On the contrary, when the value is minimum, it means the possibility of finding the pattern in that specific point is close to zero. The coefficient signal, see Figure 6.b, represents the result after applying MRAs to the walking signal. The new definition of the gesture for generating a pulse is different now. The time that a signal is descending from top to bottom defines the width of pulses. The associated function to the forward movement will be active when the pulse is active. If the difference between max and min is high, then the pulse width is longer. This mechanism creates a train of pulses with different widths (Figure 6.c).

A new mother wavelet was designed to analyze the walking signal by MRAs method. The first question which comes to our mind is "can we generalize this wavelet to a real-time process? How much is the precision in the practical test?" To answer these questions we need to analyze the performance of the adapted mother wavelet in the presence of noise, scale variation, superposition and time shift. Figure 7.a.1,a.2 shows the performance of the mother wavelet with the presence of Gaussian noise (50% of the pattern amplitude), scale variation (scale: 4, 8) and time shift (20, 40). As seen, the result is quite precise. Figure 7.b.1,b.2 demonstrates the result for the signal superposition with different scale (8, 32) factors and the presence of noise. The base triangular signal (T) and the noise are the same as in the previous test (Figure 7.b.1). The result of the test is shown in Figure 7.b.2, only very small error is created in the second test comparing to the first test, which can be neglected and does not affect the performance. The superposition of two patterns happens very rarely in gesture analysis.

Another test was carried out to show the effectiveness of the proposed wavelet compared to other most used mother wavelets. Because of the similarity, Debouche (db6) mother wavelet was selected for the test. The same settings as in the previous tests were used in the test. The comparison of the results depicted in Figure 7.c1,c2, shows that db6 creates a lot of artifacts and the precision is not comparable with the proposed method wavelet. "Haar", "bio", "db1-6" were tested too and the error was quite high with low detection precision. Figure 8 shows the experiment of the proposed gesture navigation in the test platform.

Seventeen subjects (4 females and 13 males) participated in the IVE locomotion study. They ranged 31.58 ± 12.69 years old in age and 74.65 ± 15.22kg in weight. Participants were subjected to two experiments: navigation in IVE by 1) walking 2) fly-stick. The final score of Visually Induced Motion Sickness (VIMS) is selected to evaluate these two navigation mechanisms. The VIMS score is calculated by the SSQ proposed by Kennedy [10]. The study was carried out inside a real-scale 3D model of a building while projecting in the CAVE$^{\text{TM}}$system. The statistical analysis shows that walking in place induces less VIMS than fly-stick ($F(1, 16) = 21.16, p = 0.003$).

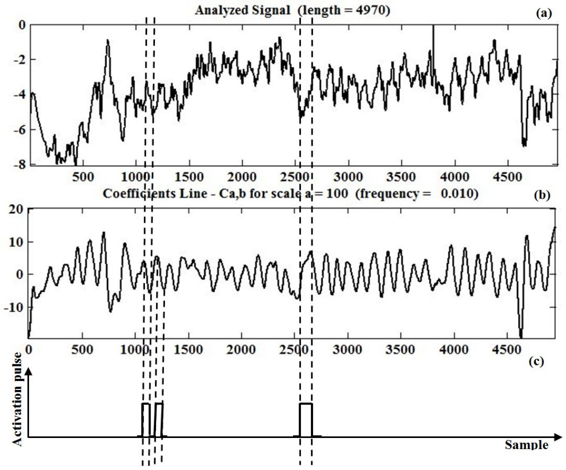

Figure 6. Generating a movement command by pattern matching.

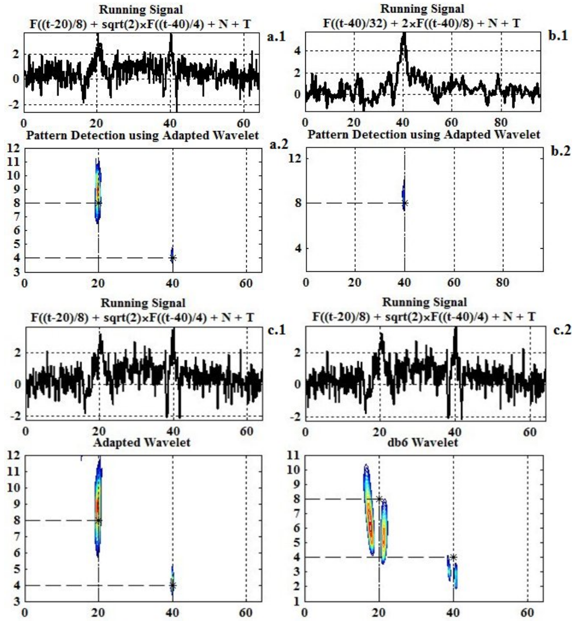

Figure 7. Performance of the mother wavelet with the presence of noise and superposition effect (a.1, a.2, b.1, b.2), comparison of the adapted (c.1) and db6 (c.2) mother wavelets.

7. Conclusion

Due to involuntary movement of the user along the Z-axis (from the sensor to the user), the amplitude of the walking signal changes from time to time and a simple threshold

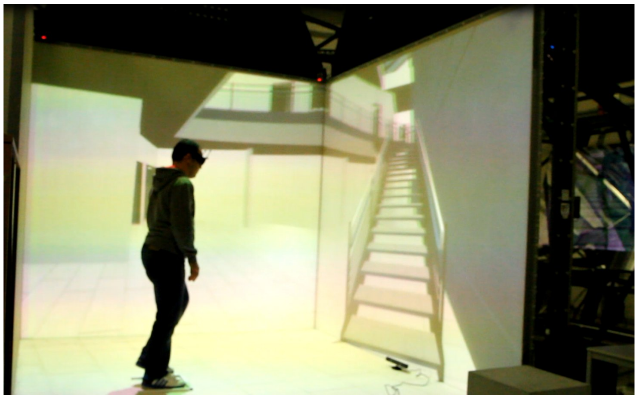

Figure 8. Practical result of the proposed method.

is not a good criterion to interpret the walking signal. On the contrary, using an adapted mother wavelet and multi-resolution analysis will help to detect the instances where a walking action happens. Consequently we can force the 3D scene to move forward/backward in those instances. Since the adapted wavelet has a very precise performance in the presence of noise, variation of the scale and superposition with other signals, we can trust the adapted mother wavelet in a similar situation.

References

[1] P. Brodal. *The central nervous system*. Oxford University Press, USA, 2010.

[2] Y.-P. Chang, D.-J. Lee, J. Moore, A. Desai, and B. Tippetts. Finger tracking for hand-held device interface using profile-matching stereo vision. In *IS&T/SPIE Electronic Imaging*, pages 86620H–86620H. International Society for Optics and Photonics, 2013.

[3] J. O. Chapa and R. M. Rao. Algorithms for designing wavelets to match a specified signal. *Signal Processing, IEEE Transactions on*, 48(12):3395–3406, 2000.

[4] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. Surround-screen projection-based virtual reality: the design and implementation of the cave. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 135–142. ACM, 1993.

[5] A. Febretti, A. Nishimoto, T. Thigpen, J. Talandis, L. Long, J. Pirtle, T. Peterka, A. Verlo, M. Brown, D. Plepys, et al. Cave2: A hybrid reality environment for immersive simulation and information analysis. In *IS&T/SPIE Electronic Imaging*, pages 864903–864903. International Society for Optics and Photonics, 2013.

[6] E. Foxlin and M. Harrington. Weartrack: A self-referenced head and hand tracker for wearable computers and portable vr. In *Wearable Computers, The Fourth International Symposium on*, pages 155–162. IEEE, 2000.

[7] P. Fuchs, G. Moreau, and P. Guitton. *Virtual reality: concepts and technologies*. CRC Press, Inc., 2011.

[8] A. Kageyama and Y. Masada. Applications and a three-dimensional desktop environment for an immersive virtual reality system. *arXiv preprint arXiv:1301.4535*, 1:1–6, 2013.

[9] A. Kar. Skeletal tracking using microsoft kinect. *Methodology*, 1:1–11, 2010.

[10] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[11] K. Khoshelham. Accuracy analysis of kinect depth data. In *ISPRS workshop laser scanning*, volume 38, page 1, 2011.

[12] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo. Kinect identity: Technology and experience. *Computer*, 44(4):94–96, 2011.

[13] M. Mackay, R. G. Fenton, and B. Benhabib. Pipeline-architecture based real-time active-vision for human-action recognition. *Journal of Intelligent & Robotic Systems*, 1:1–23, 2013.

[14] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi. Matlab wavelet toolbox user\'s guide. version 3. *Mathwork website*, 1:1–360, 2004.

[15] C. Papadopoulos, D. Sugarman, and A. Kaufmant. Nunav3d: A touch-less, body-driven interface for 3d navigation. In *Virtual Reality Workshops (VR), 2012 IEEE*, pages 67–68. IEEE, 2012.

[16] R. Radkowski and J. Oliver. A hybrid tracking solution to enhance natural interaction in marker-based augmented reality applications. In *ACHI 2013, The Sixth International Conference on Advances in Computer-Human Interactions*, pages 444–453, 2013.

[17] E. A. Suma, B. Lange, A. Rizzo, D. Krum, and M. Bolas. Faast: The flexible action and articulated skeleton toolkit. In *Virtual Reality Conference (VR), 2011 IEEE*, pages 247–248. IEEE, 2011.

[18] R. M. Taylor II, T. C. Hudson, A. Seeger, H. Weber, J. Juliano, and A. T. Helser. Vrpn: a device-independent, network-transparent vr peripheral system. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 55–61. ACM, 2001.

[19] R. I. Thompson, L. J. Storrie-Lombardi, R. J. Weymann, M. J. Rieke, G. Schneider, E. Stobie, and D. Lytle. Near-infrared camera and multi-object spectrometer observations of the hubble deep field: Observations, data reduction, and galaxy photometry. *The Astronomical Journal*, 117(1):17, 2007.

[20] M. Usoh, K. Arthur, M. C. Whitton, R. Bastos, A. Steed, M. Slater, and F. P. Brooks Jr. Walking¿ walking-in-place¿ flying, in virtual environments. In *International Conference on Computer Graphics and Interactive Techniques: Proceedings of the 26 th annual conference on Computer graphics and interactive techniques*, volume 1999, pages 359–364, 1999.

[21] R. Van Liere and J. D. Mulder. Optical tracking using projective invariant marker pattern properties. In *Virtual Reality, 2003. Proceedings. IEEE*, pages 191–198. IEEE, 2003.