Using Hand Gesture and Speech in a Multimodal Augmented Reality Environment

Conference Paper · May 2007

DOI: 10.1007/978-3-540-92865-2_18 · Source: DBLP

Using Hand Gesture and Speech in a Multimodal Augmented Reality Environment

M.S. Dias[1,2], R. Bastos[1,], J. Fernandes[1], J. Tavares[1], P. Santos[2]

[1] ADETTI Av. das Forças Armadas, Edifício ISCTE 1600-082 Lisboa (+351) 21 782 64 80, Portugal
[2] MLDC - Microsoft Language Development Center, Edifício Qualidade C1-C2, Av. Prof. Doutor Aníbal Cavaco Silva, Tagus Park, 2744-010 Porto Salvo (+351) 96 2093324, Portugal

Abstract. In this work we describe a 3D authoring tool which takes advantage of multimodal interfaces such as gestures and speech. This tool allows real-time Augmented Reality aimed to aid the tasks of interior architects and designers. This approach intends to be an alternative to traditional techniques. The main benefit of using a multi-modal based augmented reality system is the provision of a more transparent, flexible, efficient and expressive means of human-computer interaction.

Keywords: Gesture Tracking, Augmented Reality, 3D Authoring Tool, Speech, Multimodal interfaces

1 Introduction

Information Technologies (IT) use is now spread out through the world in several areas, such as medicine, communications, industry, media and many others. IT professionals require new ways of interacting with computers using more natural approaches. Computer Vision, as an example, has enabled these professionals to explore new ways for humans to interact with machines and computers [AM00] [SKL00]. The adoption of multimodal interfaces in the framework of augmented reality, is one way to address these requirements [O03]. Augmented Reality, an area that is getting recognition in the scientific arena, as evolved by improving early vision-based tracking techniques [K01], with more efficient and reliable fiducial marker tracking techniques [SSB*06][DB06], texture tracking [BD05][KTB*03], marker-less tracking [SB02][SFZ00] or even tangible artifacts manipulation [DJB*04][ZCP04]. Other authors have explored multi-modal interaction in Augmented Reality, with a special focus in human-computer interaction modalities, such as speech command and control [HBB04], gesture recognition [SML*04][MHP*06] or human motion capture [WWX*06]. Our main goal was to develop a multimodal augmented reality application that could explore the integration of gesture and speech, and could provide a valuable aid for interior design professionals in their work. It has been shown that multimodal interfaces can reduce errors and time of action, improving efficiency and effectiveness while executing a certain task. This is due to the fact that the user can freely chose the interface which is considered more appropriated to execute each one of the tasks [OV96]. A good example of the kind of interaction addressed in this paper is Bolt's "Put That There" [B80], which integrates speech recognition and gesture pointing interaction. Gestures interfaces generally involve cumbersome tethered devices and gloves based on joint angles. One alternative is by using computer vision gesture recognition techniques [ESY*03], such as multi-scale color feature detection, view based hierarchical hand models, particle filtering and other techniques. In this paper we have used O.G.R.E - Open Gesture Recognition Engine, a framework previously developed by some of the authors [DNB*06], in order to track and recognize some cases of hand gestures, which is part of the in-house MX Toolkit library [DBS*03]. O.G.R.E, requires a single video camera that captures the user's hand motion and recognizes, in real time, a set of known hand poses or other types of gestures that can be used in any type of final application that may require this type of HCI modality, offering the possibility to trigger user-specified actions, activated by different hand gestures. With O.G.R.E we can use hand gestures to move objects using trajectories and to select actions, using static poses. Although there are other viable computing environments to track hand gestures [CMV04], we have chosen this computer vision based engine, since it has been extensively tested in our group and has proven to accomplish good results, especially for static hand poses which could fit our requirements. This engine also features a reduced hardware complexity, since only needs one web cam to capture the pose of the hand and its motion. Another goal of our work was to provide "familiar" interfaces to the user, which most certainly is unfamiliar with multimodal interaction. The use of gesture and speech in an augmented reality authoring framework is the main contribution of this paper, which is organized as follows:

- In section **2. Previous Work**, we provide a short overview on related areas and prior knowledge where our research group has been working for the past years, namely in the fields of augmented reality and vision-based gesture recognition.
- In section **3. User Requirements,** we present an extended description of the main user driven goals and features of the developed system and application.
- In section **4. System Architecture,** we present the system architecture and describe how the data flow is processed between different software and hardware modules.
- In section **5. Use Case Diagram and Discussion,** we describe several functionalities of the developed application, in the context of a use case diagram and discuss some of the encountered problems during system usage.
- In section **6. System Restrictions,** we present the system limitations and possible solutions for it.
- In section **7. Conclusion and Future Work,** a conclusion about the results of our work is drawn, and some future work features and improvements are discussed.

2 Previous Work

This research group has been interested in the fields of Computer Vision, Augmented Reality and Gesture Recognition for the past nine years, therefore developing a variety of work in these areas. This paper exploits some previously obtained knowledge, namely the MX Toolkit library [DBS*03]. This library conveys a platform, which allows the programmer to combine multimodal interfaces with 3D object interaction and visualization, applied to augmented reality scenarios. It features high level functionalities, such as user tracking using ultra-sounds and gyroscopes [IS07], 3D scene graph and rendering using Open SceneGraph [OSG07], spatial audio using Open Audio Library (OpenAL) [OAL07], speech recognition using MS-SAPI [SAPI07], several tracking systems (marker, texture and 3D artifact tracking); serial communication using RS232 protocol [RS232_07]; TCP/IP communication for collaborative environments; generic video input frame grabber; and other basic functionalities like image processing using Intel's OpenCV Library [OCV07]. One of the relevant modules used in this work is O.G.R.E.. This module uses background detection and subtraction, color segmentation and contour tracking for detecting the human hand pose. The result is then compared with a pre-defined hand poses library to obtain the final recognition result. This module also includes several algorithms for testing static hand poses - discrete cosine transform (DCT), pair-wise geometrical histogram (PWGH), template matching (TM) and others. For this paper we decided to use the latter, since it's the one which presents better pose recognition results [DNB*06]. Mx Toolkit also features two modules for marker tracking: X3M [DB06], TTS [BD05] and ARToolkit [K01]; one module for planar texture tracking TTS [BD05]; and one module for 3D object tracking - ARTIC [DJB*04]. The X3M technique, similar to ARToolkit, is a marker tracking module based on fiducials markers with a planar squared topology and high contrasted contours, which allows tracking of full colored natural textures at high processing speed. The TTS module features an algorithm for texture tracking, based on natural features extraction and template matching. The method is oriented to planar objects with arbitrary textures, but with rectangular topology and well contrasted contours. This module does not require any fiducial marker for the first pose extraction, in contrast to other similar methods [KTB*03], and combines methods like Direct Linear Transformation [AK71], Kalman filtering [K60] and Template Matching. Since the final goal of this paper was the creation of an augmented reality computational application that is able to help interior design professionals in an augmented reality authoring environment, we have integrated a previous developed Augmented Reality Authoring tool, based in MX Toolkit, the Plaza [S05]. Plaza is a 3D AR authoring module that allows the user to manipulate and modify 3D objects loaded from a predefined database either in a VR environment, in an AR scenario or in both.

3 User Requirements

Traditionally interior design has been done using the "paper and pencil" metaphor. However, this method is slow for the designer and sometimes hard to communicate for the client. Recently, a set of computational applications for interior design through the use of 3D models, both for rooms and for furniture elements, has become standard in the industry and retail. If on one hand these tools make the design task easier, on the other hand the creation of 3D models is still time consuming. The inclusion of furniture elements into scenes implies searching large databases. Another problem is the visualization of results. Although the presentation of the solution is better than before, these systems continue to show a small, limited and hard to navigate virtual model, without the notion of the real dimensions. Our work envisages the creation of a tool for architects and interior designers which allows, via multimodal interaction (gesture and speech), the designers or the clients, to visualize the implementation of real size furniture using augmented reality. The tool has to be capable of importing, disposing, moving and rotating virtual furniture objects in a real scenario. The users should be able to take control of all actions with gestures and speech, and should be able to walk into the augmented scene, seeing it from a variety of angles and distances. The tool has also to address a common problem in this kind of authoring tools, which is the large variety of objects available, making the navigation and selection of the desired object difficult. One of our primary objectives was also to create "familiar" interfaces so that no training to use this software was needed, making it attractive to everyone.

4 System Architecture

The proposed logical architecture of the system is depicted in Fig. 1 and can be divided in two modules: Plaza, responsible for Augmented Reality authoring and Speech Recognition and the Gesture Recognition Server, responsible for Hand Gesture recognition. Both modules use the MX Toolkit library and communicate through the TCP/IP COM module. The Gesture Recognition Server also maintains a Gesture Database, which will be used at runtime for gesture matching.

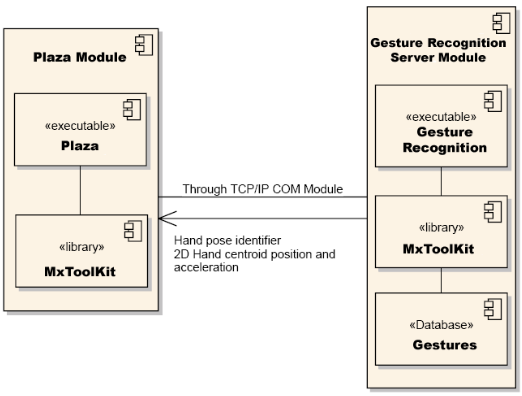


Fig. 1. – System Architecture Diagram.

The MX Toolkit (Fig. 2) has 14 different modules, but only 8 are used in this work. The system libraries and operating system used are as follows: Microsoft Windows XP; DirectX SDK v 8.1 (or higher); OpenGL; INTEL OpenCV; MX Toolkit; Microsoft Speech SDK 5.1.

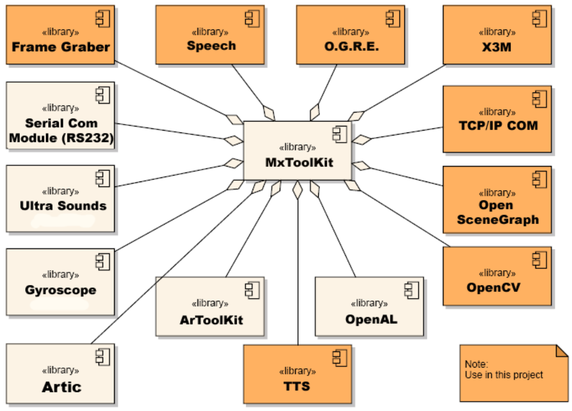


Fig. 2. – MX Toolkit library components (only dark shaded ones are used in this work).

The system hardware (Fig. 3) or deployment architecture can be also divided in two distinct modules: Gesture Recognition Server and Plaza. The Gesture Recognition Server is connected to a fixed webcam device. The Plaza module is connected to a wireless webcam device and to a video see-through glasses device. Both modules have Wireless 802.11g interfaces and communicate using TCP/IP protocol. The fixed webcam device is only used for gestures recognition, while the wireless webcam provides an augmented reality image of the room, which can be perceived in the video see through glasses. The hardware configuration used is as follows: Two laptops with the following minimum configuration - Pentium III processor (recommended is a Pentium Dual Core processor); 512Mb of RAM (recommended is 1024 Mb); Geforce2 with 32Mb of video memory (recommended Geforce7200 with 128Mb of video memory); Wireless 802.11g adapter; Webcam with 320x240 resolution, working at 30 Hz and manual white balance control; video-see-through glasses.

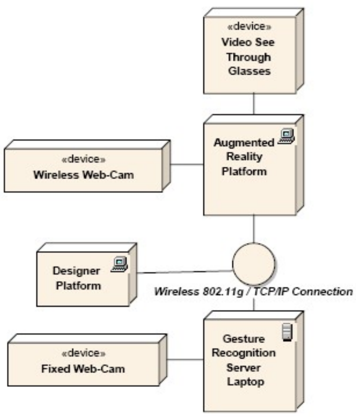


Fig. 3. - Hardware block diagram.

The processes sequence and data flows of the system are straightforward:

1. Launch Gesture Recognition Server.
2. Launch Plaza. It launches also Speech Recognition and selects the applicable language.
3. Plaza registers the gestures for notification, sending information to the Gesture Server.
4. At each frame, the Gesture Server sends Plaza relevant information (in a non null case), about the detected gesture (hand pose identifier) and the hand´s position (2D centroid coordinates) and acceleration.
5. Plaza, at each frame, receives a gesture identifier (in a non null case), as well as other gesture information (position, acceleration).
6. Plaza uses the received information as multimodal interface and authors the Augmented Reality environment.

The user´s gesture interaction is performed using a fixed camera near the Gesture Recognition Server (see Fig.4).

Fig. 4. –Example of user interaction through gestures and voice

5 Use Case and Discussion

In order to associate commands to hand gestures, a predefined gesture database must be created. The system supports several user profiles that can be created using the Gesture Recognition Server, either in online or offline modes. Each profile corresponds to different gesture templates, since segmentation conditions may vary from user to user (skin tone, hand size, etc.). For each command there is an associated hand pose gesture. A map between gestures and commands used by Plaza is shown below (Fig. 5). Each one of these commands can be also invoked using the Speech Recognition Module, which is always running in background performing speech recognition. This ensures the possibility for the user to choose between issuing a command from gesture or speech. Every time the user activates an action (being a hand pose gesture or a speech command), this action is associated with a timestamp, so the system can synchronize the received instructions from the two interfaces, speech and gesture. This way the user can choose freely between the two interfaces.

At runtime, whenever the user performs a gesture or pronounces an utterance associated with a command, that command is invoked and a note will appear on the left-top of the image enumerating the current task.

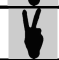
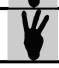
Action	Speech	Gesture
Rotate	"Rodar"	
Translate	"Mover"	
Right	"Direita"	
Left	"Esquerda"	
Up	"Cima"	
Down	"Baixo"	
Back	"Trás"	
Forward	"Frente"	
OK	"OK"	
Slow	"Lento"	
Normal	"Normal"	
Fast	"Rápido"	

Fig. 5. –Gesture / Voice database and commands

Every time a valid gesture identifier is sent to Plaza from the Gesture Recognition Server, the user will view, in the Video See-Through rendered augmented reality image, some type of feedback (circles, point paths, text, etc.) depending on the performed action. After selecting a marker to work on, the user can select any 3D object from the objects database, in order to be registered and rendered in the current augmented reality scene. To do so, the user may select the desired object by choosing the object from the Plaza interface. The user can perform two types of transformations on the object: translate, or rotate. Each one of these operations may be activated by a gesture or a speech command. The process of these operations is similar. When the user selects the transformation, the directions available will show on screen. He/She must then choose one direction by using a voice or gesture command. Then the user will be capable of applying the desired actions on the object and confirm it by saying "Ok" or using a gesture. The user must interact with this interface using the same method proposed when selecting a 3D object or simple by saying "direita[1]" or "esquerda[2]".
In case of "OK", all transformations will be saved and the system will be in standby, waiting for another command. The user can repeat this process until a satisfactory

1 "Right"
2 "Left"

result is achieved, by applying new transformations to the selected object. Since this work uses the Plaza module, all these mentioned operations and other (light positioning, texture mapping, clipping plane manipulation, material selection, etc.) can be executed directly on the Plaza Application GUI using common HCI interfaces (keyboard, mouse and space-mouse), which may aid in the initial authoring of the AR scene.

To determine the usability of the system we developed a test, based on simple geometric transformations applied to virtual models using gesturing and voice interfaces. Our objective was to determine which interface the user would prefer and why. We applied the same test 3 times to each subject, using each time a different interface (Voice, Gestures and the two together). The usability experiment was run on 11 unpaid users, which were students in their early 20's from the undergraduate course of ISCTE, Instituto Superior de Ciências do Trabalho e da Empresa in Lisbon, Portugal and some developers with computer graphics experience from our laboratory (see Fig. 8).

First, a 10 minute briefing, was given to each tester, explaining the project and features, showing what could be done with the system. After this briefing, each subject was given 2 minutes to play freely with the system and expose any doubt. Then the users performed the 3 tests, first using only voice, then using only gestures and finally using the two interfaces freely. After all the tests, each subject was given 5 minutes to fill a questionnaire related to the tests and their feedback. It was revealed that the use of the two interfaces was definitively the best way to interact with the system reducing in 20% (average) the time to complete the test. Most of the users had used voice to activate simple commands like "Mover[3]" and "Rodar[4]" or to modify the speed (see Fig. 6 and Fig. 7)). The use of gestures was preferred to move the objects, after the corresponding action has been selected using speech. The subjects provided valuable feedback during and after the experiment. Several subjects said that the three velocity factors should be refined, because there was a much bigger difference between fast and normal than the difference between normal and slow.

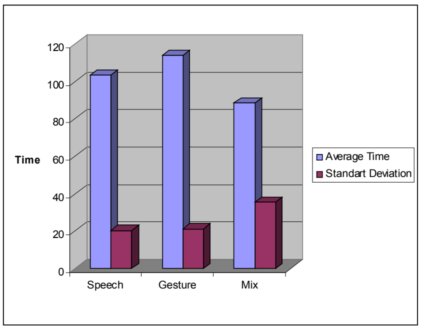


Fig. 6. – Average time for the test (time unit in seconds)

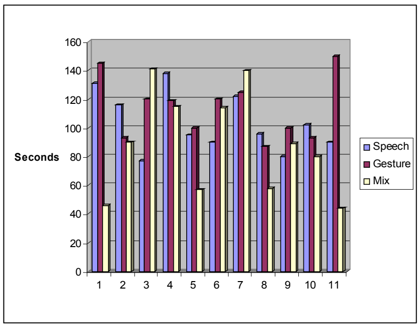


Fig. 7. – Time each subject used for the test

3 "Move"

4 "Rotate"

Fig. 8. – Some of the Usability Test subjects

6 System Restrictions

The system shows some difficulties when tracking gestures, in the presence of bad lighting conditions. The shadows turn color segmentation inaccurate, therefore gesture matching may fail. Another limitation of our system is the lack of mobility in what concerns to the camera that is tracking the gestures. This happens as we are using background subtraction and any camera movement would change the background template. When this occurs, the Gesture Recognition Server needs to recalibrate the background, interrupting the authoring process. Since the camera calibration process is very sensitive to movement, it would be very difficult to maintain the camera steady if it was attached to the user, or some movable object. The speech recognition system is robust but requires the use of a headset to maximize the signal to noise ratio. Another issue is that the user mobility is restrained by the Video-See-Trough glasses cable length. A solution for this problem would be to use a wireless video transceiver and a mobile power source (battery). An additional problem arises, since we don't know the room geometry and, in certain conditions, we cannot calculate the occlusions of the virtual model as it stands behind of a real obstacle.

7 Conclusions and Future Work

In this paper we have presented an application developed to help interior design professionals using multimodal interaction in augmented reality. The HCI modalities used are hand gestures and speech commands. The system also combines different tracking systems in order to obtain the virtual camera pose. We have adopted a client/server topology, based on Wireless 802.11g TCP/IP connections, since the gesture recognition module must be separated of the authoring module, due to efficiency restraints. The created system is still a preliminary prototype, with some restrictions, and is the result of some years of research in augmented reality and computer vision areas. The application works in real-time and is able to detect and track static hand poses and hand movements, using them to control and manipulate all objects in the scene. In the near future, we will integrate some other technologies on this system, in order to help the performance and usability. Using multi-markers, a technique present in X3M, we will be able to associate the same object to several markers, and this way, if one of the markers is occluded , the system will continue to track the object, as long as one of the markers remains visible. We also would like to integrate the ARTIC module, as another multimodal interface, giving the user other alternatives to interact with the system and expanding the way tasks are accomplished. Searching a 3D object database for an item, can be difficult and inefficient. To solve this issue, we want to implement a system based on the CaLi [JF00] library. This library is based on fuzzy logic and it was developed for recognizing drawn forms in calligraphic interfaces. This way, we will allow the user to draw a sketch of a desired object using hand written gestures, and then the algorithm will search for a related shape on the database and will sort by similarity. To improve the usability of the system, we want to elaborate more complex usability evaluation tests on several test subjects, so we can take conclusions about future improvements of this platform.

8 REFERENCES

[AK71] Abdel-Aziz, Y.I. and Karara, H.M., "Direct Linear Transformation into Object Space Coordinates in Close-Range Photogrammetry", in Procedures of Symposium of Close-Range Photogrammetry, pages 1-18, January 1971.

[AM00] Abowd, G.D., Mynatt, E.D., "Charting Past, Present, and Future Research in Ubiquitous Computing", in ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 7, Issue 1, pages 29-58, March 2000.

[B80] Bolt, R.A., "Voice and Gesture at the Graphics Interface." ACM Computer Graphics, 14.3 pages 262-270, 1980.

[BD05] Bastos, R., Dias, J.M.S., "Fully Automated Texture Tracking Based on Natural Features Extraction and Template Matching", in ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, Valencia, Spain, 2005.

[CMV04] A. Camurri, A., Mazzarino, B, Volpe, G. "Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library, in A. Camurri, G. Volpe (Eds.), "Gesture-based Communication in Human-Computer Interaction", LNAI 2915, Springer Verlag, 2004.

[DB06] Dias J.M.S., Bastos, R., "An Optimized Marker Tracking System", in Eurographics Symposium on Virtual Environments pages 1-4, 2006.

[DBS*03] Dias, J. M.S., Bastos, R., Santos, P., Monteiro, L., Silvestre, R., "Developing and Authoring Mixed Reality with MX Toolkit", in The Second IEEE International Augmented Reality Toolkit Workshop ART 03, Toki, Japan, 2003.

[DJB*04] Dias, J. M. S., Jamal N., Bastos R., Silva P., "ARTIC: Augmented Reality Tangible Interface by Color Evaluation", in Interacção 2004, Faculdade de Ciências da Universidade de Lisboa, Lisbon, July 2004.

[DNB*06] Dias, J. M. S., Nande, P., Barata, N., Correia, N., "O.G.R.E. – Open Gestures Recognition Engine, a Platform for Gesture-based Communication and Interaction", in Human-Computer Interaction and Simulation, 6th International Gesture Workshop, GW 2005, Ile de Berder, France, May 18-20, 2005, Revised Selected Papers, Sylvie Gibet, Nicolas Courty, Jean-Francois Kamp (Eds.), Lecture Notes in Artificial Intelligence, Vol. 3881, Springer-Verlag 2006, XIII, ISBN 3-540-32624-3, March 2006

[ESY*03] Elgammal, A., Shet, V., Yacoob, Y., Davis, L., "Learning dynamics for exemplar-based gesture recognition" in IEEE Conference on Computer Vision and Pattern Recognition, pages 571-578, 2003.

[HBB04] Heidemann, G., Bax, I., Bekel, H., "Multimodal interaction in an augmented reality scenario", in Proceedings of the 6th international conference on Multimodal interfaces, pages 53-60, 2004.

[IS07] Intersense: http://www.intersense.com, visited at 15.01.2007.

[JF00] Jorge, J., A., Fonseca, M., J., "Using Fuzzy Logic to Recognize Gestures Interactively", In FUZZ-IEEE'2000, Intl. Conference on Fuzzy Systems, volume 1941, pages 265–275, SanAntonio, Texas, USA, May 2000.
[K01] Kato, H., "Developing Applications with ARToolkit", SIGGRAPH 2001 Course Notes 27, 2001.
[K60] Kalman, R. E., "A new approach to linear filtering and prediction problems", in Trans ASME, vol. 82, no. D, pages 35-45, 1960.
[KHW95] D. Kessler, L. Hodges, N. Walker, "Evaluation of the Cyber Glove as a Whole-Hand Input Device." In ACM TOCHI, pages 263-283, 1995.
[KTB*03] Kato, H., Tachibana, K., Billinghurst, M., Grafe, M., "A Registration Method based on Texture Tracking using ARToolKit", in The Second IEEE International Augmented Reality Toolkit Workshop, Nishi-Waseda Campus, Waseda University, Tokyo, Japan, October 2003.
[MHP*06] Morris, M., R., Huang, A., Paepcke, A., Winograd, T., "Cooperative gestures: multi-user gestural interactions for co-located groupware", in Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 1201-1210, 2006.
[O03] Oviatt, S.L., "Multimodal interfaces", in The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, chap.14, pages 286-304, J. JACKO AND A. SEARS, Eds. Lawrence Erlbaum Assoc., Mahwah, NJ, 2003.
[OAL07] Open Audio Library: http://www.openal.org/, visited at 15.01.2007.
[OCV07] Intel Open Computer Vision: http://sourceforge.net/projects/opencvlibrary
[OSG07] Open Scene Graph: http://www.openscenegraph.org/, visited at 15.01.2007.
[OV96] Oviatt, S.L., vanGent, R.. "Error Resolution During Multimodal Human-Computer Interaction", in Proceedings of the International Conference on Spoken Language Processing, Vol. 2, pages 204-207, University of Delaware Press, 1996.
[RS232_07] RS232 Standard: http://www.camiresearch.com/Data_Com_Basics/RS232_standard.html
[S05] Santos, P., "AR Authoring for 3D Design", in Master of Science Thesis in Computers and Telecommunications Engineering, ADETTI, Portugal, 2005.
[SAPI07] Microsoft Speech Library: http://www.microsoft.com/speech/default.mspx
[SB02] Simon, G., Berger, M-O., "Reconstructing while registering: a novel approach for markerless augmented reality", in ISMAR' 02 Conference Proceeding, 2002.
[SFZ00] Simon, G., Fitzgibbon, A., Zisserman, A., "Markerless Tracking using Planar Structures in the Scene", in ISMAR' 00 Conference Proceeding, 2000.
[SKL00] Segen, J., Kumar, S., "Look Ma, No Mouse! – Simplying human-computer interaction by using hand gestures", in Communications of the ACM, Vol. 43, Issue 7, pages 102-109, July 2000.
[SML*04] Störring, M., Moeslund , T.B., Liu, Y., Granum, E., "Computer Vision-Based Gesture Recognition for an Augmented Reality Interface", in 4th IASTED International Conference on VISUALIZATION, IMAGING, AND IMAGE PROCESSING, pages 766-771, Marbella, Spain, September 2004.
[SSB*06] Santos, P., Stork, A., Buaes, A., Jorge, J., "PTrack: Introducing a Novel Iterative Geometric Pose Estimation for a Marker-based Single Camera Tracking System", in Proceedings of the IEEE Virtual Reality Conference (VR 2006) - Volume 00, page 20, 2006.
[WWX*06] Wen, G., Wang, Z., Xia, S., Zhu, D., "From motion capture data to character animation", in Proceedings of the ACM symposium on Virtual reality software and technology, pages 165-168, 2006.
[ZCP04] Zhou, Z., Cheok, A., Pan, J., Li, Y., "Magic Story Cube: an interactive tangible interface for storytelling", in Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology, pages 364-365, 2004.