EG 2005 Tutorial on Mixed Realities in Inhabited Worlds

Organizer: Nadia Magnenat-Thalmann

Presenters: Nadia Magnenat-Thalmann (University of Geneva), Daniel Thalmann (EPFL), Pascal Fua (EPFL), Frederic Vexo (EPFL), HyungSeok Kim (University of Geneva)

Institutions: University of Geneva, Geneva, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Address: MIRALab,
24, rue du General Dufour
1211 Geneve 4 Switzerland

e-mail: thalmann@miralab.unige.ch

Phone: +41 22 379 77 69

Fax: +41 22 379 77 80

URL: http://www.miralab.unige.ch

Keywords: I.3.7 [Three-Dimensional Graphics and Realism]: Believability, Virtual Reality, Augmented Reality, Virtual Humans, Mixed Reality, Real-time, Presence, Emotion

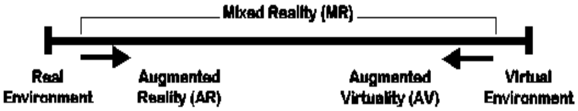
Necessary background and potential target audience for the tutorial:
experience in virtual reality is recommended but not mandatory.

1. Outline of the tutorial

1.1 Concepts and State of the Art of mixed realities in inhabited worlds

1.1.1 Mixed Realities in inhabited worlds

Mixed Reality covering the spectrum from Reality to Virtual Reality. We will emphasize Augmented Reality which augments the user's view of the real world by composing 3D virtual objects with their real world counterparts, necessitating that the user maintains a sense of presence in that world. We also present the concept of Augmented Virtuality where the real part is much less important. We then focus on Virtual Inhabited Worlds.



After a State-of-the-Art of the main technologies and concepts, we will survey a few applications in medicine, psychiatry, tourism, and entertainment.

1.1.2 Believability and Presence

In the Mixed Reality (MR) environment, the concept of cyber-real space interplay invokes such inter-affective experiences that promote new patterns of believability and presence. Believability is a term used to measure the realism of interaction in the MR environments. Presence is defined as the measure that is used to convey the feeling of 'being there'. In this session, a concept of Believability and Presence for MR is presented, starting from the sensory level interaction to the perceptual level interaction, focused on the inhabited environment. We show an example of inhabited MR environment which show strengthened presence but without enough believability, due to limited interaction between the real participants and the virtual characters. Among present technologies for MR environment, we introduce a new concept of 'affective registration' to enhance sensory level experience through keeping consistency of the virtual environment with the real environment. We also argue that future steps in MR enabling technologies should cater for enhanced social awareness of the virtual humans to the real world and new channels for interactivity between the real users and virtual actors. Only then the believability of interactions in a MR environment will be enhanced and allow for compelling real experiences.

1.2 Perception, Sensors and Immersive hardware for MR in Inhabited Worlds

1.2.1 Vision Based 3D Tracking and Pose Estimation for MR

In augmented reality applications, tracking and registration of cameras and objects are required because, to combine real and rendered scenes, we must project synthetic models at the right location in real images. It is therefore crucial to seamlessly integrate tracking and object detection algorithms into MR systems.

We will focus on Computer Vision based approaches because they have the potential to yield non-invasive, accurate and low-cost solutions to this problem, provided that one is willing to invest the effort required to develop sufficiently robust algorithms. In some cases, it is acceptable to add fiducials, such as LEDs or special markers, to the target objects to make the task easier. However, since this is not always an option, it is much more desirable to rely on naturally present features, such as edges, corners, or texture. This makes the problem far more challenging but is worthwhile because it yields much more versatile systems. We will first introduce the key mathematical tools required for 3D tracking. We will then present marker-based techniques that use either point fiducials or planar markers sometimes used to make the tracking task easier. Next we will focus on techniques that rely on natural features for both tracking and detection. Finally, we will present a specific application and discuss potential future developments.

1.2.2 Perception and sensors for Virtual Humans

A Virtual Human is situated in a Virtual Environment (VE) equipped with sensors for vision, audition and tactile, informing it of the external VE and its internal state, it may also be aware of the user through real sensors like cameras and microphones. A Virtual Human possesses effectors, which allow it to exert an influence on the VE and a control architecture, which coordinates its perceptions and actions. In order to select the appropriate actions of an actor, the behavioral module needs to know the state of the environment of the actor. However, an actor is not passive, but performs actions which might involve objects, other actors, or even the user. Moreover, the actions of an actor may cause some events. Therefore, perception is decomposed into three categories: perception of objects and actors, actions of actors and events. Perception of events is slightly more complex because events themselves are decomposed into three classes: desirable events, events happening to another actor and potential events which may or may not occur. The perception of the nature and the characteristics of an object, an actor or an action is not easily done from their 3D representation. Recognizing an action through motion is difficult as well. The adopted solution is to categorize every object, actor and action based on its nature and characteristics. We will then study the problem in the context of a group and even a large crowd.

1.2.3 Hardware for mixed reality inhabited virtual world

Of course, creation of mixed reality inhabited virtual world is a very challenging application on the software side. In fact it’s necessary to animate in real-time the complex synthetic world including its inhabitants. Nevertheless, it’s also challenging on the hardware side, because it’s needed to be able to display in real-time the 3D content on the top of the video with the help of wearable devices. Thanks to the miniaturization of the electronics, it’s now reasonable to speak about wearable devices which will help the creation of the interface for mixed reality.

This part of the course will present the new devices available such as miniatures head mounted display, wearable computer with graphic acceleration (new generation of handles devices or 3G mobile phones), head and gaze tracking or light weight geo-localization if we are developing mixed reality for outdoor purpose. Last but not least, to create mixed reality we need to receive video data from the real-world that could be just in front of us or far away in case of robot Tele-Operation. For this purpose good quality and light weight cameras are needed, we will present in this tutorial the state of art in term of light weight controllable cameras.

The course will be illustrated with some examples coming directly from the last research result such as UAV control, social phobia treatment with the help of mixed reality.

1.2.4 Emotional and conversational virtual humans

As 3D graphic techniques have matured, we are now able to create realistic 3D characters that can move and talk in real-time MRs. Linking these characters with perceptive technologies (speech recognition, vision) and interactive technologies (agents, dialogue management) is an ongoing research effort to create an Embodied Conversational Agent or ECA. An important aspect of this research is to allow ECAs to become emotional individuals. By populating the MRs with Virtual Individuals instead of animatable 3D models, users will feel more present since they are surrounded by characters that move and behave consistently, according to their personality. This can be achieved by using personality and emotion simulation models. Such models will influence the ECA behavior on all levels (perception, interaction and expression). In this tutorial, we will focus on that adaptation of the expression depending on the personality and emotional state. Facial animation methods will be presented that show how to express emotions together with speech in a generic way. Also, controlling body motion and posture using different personality and emotion parameters will be discussed.

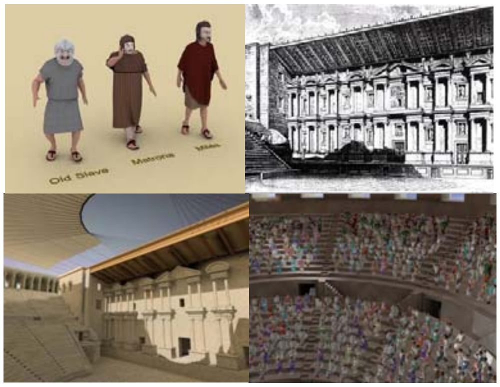
1.3 MR in various applications

1.3.1 Simulating Life in mixed realities Pompei world

The European project LIFEPLUS ‘brought to life’ the Ancient Pompeian fresco paintings, through 3D animation of their content, superimposed on their real environment. The whole experience allows the user to visit on-site, by means of an immersive, mobile Augmented Reality-based Guide featuring wearable computing and multi-modal interaction. The LIFEPLUS mobile system is required to operate in two main modes: The “sight-seeing” operational mode is designed to support the visitor with location based multimedia information facilitating sight-seeing of the area by provision of both practical and historical information in form of text, images, short movies overlaid on the head mounted display. In AR simulation mode, the visitor is exposed to the VR simulation scenario blended into the real imagery of the site.

1.3.2 Simulating actors and audiences in ancient theaters

Ancient theatres were places where a large amount of people from different extractions could gather for important social events such as, but not limited to, representations of comedies or tragedies in which male actors, musicians and dancers wearing masks, in accordance to their specific role and character, would perform on a scene in front of the public. In this case-study we will present the ancient theatre of Aspendos ad the odeon of Aphrodisias, located in Turkey near the city of Antalya, virtually restituted in a real-time 3D inhabited environment as Roman buildings of the third century. The essential steps that are to be considered, to stage a real-time virtual reenactment of a Roman play will be illustrated: we both focus our attention on the creation of 3D virtual actors capable of performing a selected Roman play and on the creation of a virtual audience that is to emotionally respond to the events that are occurring on stage. We survey historical sources concerning the appearances of the Romans (such as clothes, shoes, hairstyles and bodies), the distribution and behaviors of spectators in ancient Roman times and the architecture.



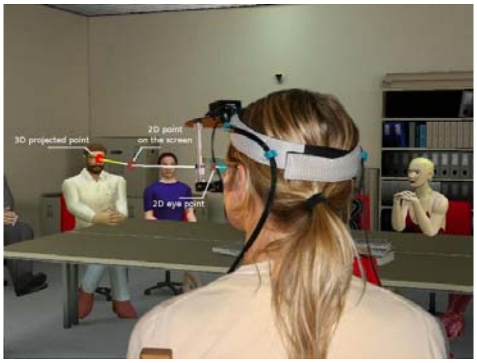
Then we present the different approaches and techniques that are needed in order to create believable virtual embodiments, that are meeting the inherent limitations of a real time virtual environment in terms of the trade-off between the precision of the simulation and the achieved frame-rate, and that can display, on the other hand, all their necessary features, such as cloth simulation, emotional response and proper behavior, facial animation, realistic movements and gestures.

1.3.3 MR in STAR, an industrial project

STAR is mixing images of real environments, virtual objects and virtual humans to produce mixed reality animations of an existing environment. It strives to provide factories with new digital tools to facilitate the planning and carrying out of routine maintenance inspections and revamp procedures. Such tools also contribute to put together easy-to-understand procedure manuals. STAR focuses on enhancing the quality of communications in the workplace between individuals or small groups, by creating mixed-reality animations of the procedures. Using mixed reality techniques instead of virtual reality techniques, the quality of the animations can be made to approach that of documentaries with much higher production costs (camera crew, editor, etc.). Also mixing real parts of a scene with virtual ones is more cost-effective than the creation and rendering of wholly virtual ones.

1.3.4 Feeling presence in the treatment of social phobia

The necessity to mixed reality appears recently for patient treatment. Virtual reality entered indeed the mental health field some years ago to determine if virtual reality exposure (VRE) could constitute an alternative to standard in vivo exposure for a wide spectrum of phobias. Several studies reported VRE to be as effective as in vivo exposure in the context of anxiety and behavioral-avoidance. Applications of VRE have been developed as a tool for mental health therapists for claustrophobia and acrophobia and have been successful in reducing fear of heights. In our case, we applied VRE to the case of social anxiety disorder: Social anxiety can be induced by a virtual audience and the degree of anxiety experienced is directly related to the type of virtual audience feedback received by the speaker. We can easily changed the type and the behaviour of the virtual audience (age, sex, emotions, attitude etc.) We can also analyze the behaviour of the patient using eye tracking.



2. Syllabus

1) Concepts and State of the Art of mixed realities in inhabited worlds (45 minutes)

- ➢ Mixed Realities in inhabited worlds (Daniel Thalmann)
  - Definition of Mixed Reality
  - Inhabited virtual Worlds
  - State-of-the-Art
  - Applications of Mixed Reality
- ➢ Believability and Presence (Nadia Magnenat-Thalmann and HyungSeok Kim):
  - Concept of Presence and Believability
  - Sensory level techniques to enhance presence and believability
  - Affective registration: an example of illumination registration
  - Perceptual level Presence and Believability: Introducing the interactive future

2) Perception, Sensors and Immersive hardware for MR in Inhabited Worlds (4 hours)

- ➢ Emotional and conversational virtual humans (Nadia Magnenat-Thalmann)
  - Overview Embodied Conversational Agents
  - Personality and Emotion Simulation
  - Dialogue Systems
  - Facial Animation
  - Body Animation
- ➢ Vision Based 3D Tracking and Pose Estimation for MR (Pascal Fua)
- ➢ Perception and sensors for Virtual Humans (Daniel Thalmann)
  - Concepts of virtual and real sensors
  - Virtual vision, audition and tactile
  - Perception of objects
  - Perception of real and virtual humans
  - Perception of actions and events
  - Perception and sensors for groups and crowds
- ➢ Immersive hardware needed for Mixed Realities (Frédéric Vexo)
  - Semi Immersive Large Screen
  - Miniature display solutions
  - Head Mounted Display and Head Mounted see-through Display
  - Wearable Computer
  - Wireless Data Transmission devices
  - Head and Gaze Tracking
  - 3D graphics board for mobile devices
  - Low weight Geo-localization devices
  - Miniature and Control Video Cameras

3) Case studies: MR in various applications (all speakers, 1 hour)

- ➢ Simulating actors and audiences in ancient theaters (Daniel Thalmann, Nadia Magnenat-Thalmann)
- ➢ Simulating Life in mixed realities Pompei world (Nadia Magnenat-Thalmann)
- ➢ MR in an industrial project (Pascal Fua)

- ➢ Feeling presence in the treatment of social phobia (Daniel thalmann and Frederic Vexo)

4) Conclusions and Further Discussion (all speakers, 15 minutes)

3. Resume of the presenters

Pascal Fua received a degree from Ecole Polytechnique, Paris, in 1984 and a Ph.D. in Computer Science from the University of Orsay in 1989. He joined EPFL in 1996 where he is now a. Before that, he worked at SRI International and at INRIA Sophia-Antipolis as a computer scientist. His research interests include human body modeling from images, optimization-based techniques for image analysis and synthesis, and using information theory in the area of model-based vision. He has (co)authored over 100 publications in refereed journals and conferences. He is a member of the editorial board of the IEEE journal Transactions for Pattern Analysis and Machine Intelligence and has been a program committee member of several major vision conferences.

Dr. HyungSeok Kim is a senior research assistant at MIRALab, University of Geneva. He received his PhD in Computer Science in February 2003 at VRLab, KAIST : "Multiresolution model generation of texture-geometry for the real-time rendering". His main research field is Real-time Interaction in the Virtual Environments, more specifically multiresolution modeling of shape and texture and multimodal interaction mechanisms. He has been actively participated in several European Project focused on topics of shape modeling, multimodal interaction and evoking believable experiences in the virtual environment.

Nadia Magnenat-Thalmann has pioneered research into virtual humans over the last 25 years. She obtained several Bachelor's and Master's degrees in various disciplines (Psychology, Biology and Chemistry) and a PhD in Quantum Physics from the University of Geneva. From 1977 to 1989, she was a Professor at the University of Montreal and led the research lab MIRALab in Canada. She moved to the University of Geneva in 1989, where she founded the Swiss MIRALab, an internationally interdisciplinary lab composed of about 30 researchers. She is author and coauthor of a very high number of research papers and books in the field of modeling virtual humans, interacting with them and in augmented life. She has received several scientific and artistic awards for her work, mainly on the Virtual Marylin and the film RENDEZ-VOUS A MONTREAL. She has directed and produced several films and real-time mixed reality shows, among the latest are the UTOPIANS (2001), DREAMS OF A MANNEQUIN (2003) and THE AUGMENTED LIFE IN POMPEII (2004). She is editor-in-chief of the Visual Computer Journal published by Springer Verlag and coeditor-in-chief of the Computer Animation & Virtual Worlds journal published by John Wiley.

Daniel Thalmann is Professor and Director of The Virtual Reality Lab (VRlab) at EPFL, Switzerland. He is a pioneer in research on Virtual Humans. His current research interests include Real-time Virtual Humans in Virtual Reality, Networked Virtual Environments, Artificial Life, and Multimedia. Daniel Thalmann has been Professor at The University of Montreal. He is coeditor-in-chief of the Journal of Visualization and Computer Animation, and member of the editorial board of the Visual Computer and 3 other journals. Daniel Thalmann was Program Chair of several conferences including IEEE VR 2000. He has also organized 4 courses at SIGGRAPH on human animation. Daniel Thalmann was the initiator of the Eurographics working group on Animation and Simulation which he cochaired during more than 10 years. Daniel Thalmann has published more than 250 papers in Graphics, Animation, and Virtual Reality. He is coeditor of 30 books, and coauthor of several books including the recent book on "Avatars in Networked Virtual Environments", published by John Wiley and Sons. He received his PhD in Computer Science in 1977 from the University of Geneva and an Honorary Doctorate (Honoris Causa) from University Paul-Sabatier in Toulouse, France, in 2003.

Frederic Vexo is currently senior researcher and project leader at the Virtual Reality Laboratory at the Ecole Polytechnique Federale de Lausanne, Switzerland. His research interests are multidisciplinary and include Human to Inhabited Virtual World, Haptic Interfaces, Tele Operated System, Multimodal Adaptive Interface for virtual worlds, Semantic Virtual Environment and new services for graphic mobile devices. He is author of several papers in journals and international conferences in the fields of Human Computer Interaction, Robotics and Computer Graphics. He is member of several conference program committees (AAMAS 2005, SMI 2005, SVE 2005 and IWVR2005) and expert for different institution and companies. He received his PhD in Computer Science in 2000 from University of Reims and Bachelor's and Master's degrees from University of METZ. He also has contributed to various European projects.

4. Selected Publications

M. Gutierrez, D. Thalmann, F. Vexo, Semantic Virtual Environments with Adaptive Multimodal Interfaces, 11th International Conference on Multimedia Modelling, MMM2005, Melbourne, Australia, 12-14 Jan 2005, pages 277-283

B. Herbelin, F. Riquier, F. Vexo, D. Thalmann, Virtual Reality in Cognitive Behavioral Therapy : a preliminary study on Social Anxiety Disorder, 8th International Conference on Virtual Systems and Multimedia, VSMM2002

T. Conde, D. Thalmann, An Artificial Life Environment for Autonomous Virtual Agents with Multi-sensorial and Multi-perceptive Features, Computer Animation and Virtual Worlds, Vol.15, No3-4, 2004, pp.311-318.

D. Thalmann, Control and Autonomy for Intelligent Virtual Agent Behaviour, Method and Applications of Artificial Intelligence, Lecture Notes in Artificial Intelligence 3025, 2004, pp.515-524

V. Lepetit, J. Pilet, and P. Fua. Point Matching as a Classification Problem for Fast and Robust Object Pose Estimation. In Conference on Computer Vision and Pattern Recognition, Washington, DC, June 2004.

L. Vacchetti, V. Lepetit, and P. Fua. Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking. In International Symposium on Mixed and Augmented Reality, Arlington, VA, November 2004.

A. Egges, S. Kshirsagar, N. Magnenat-Thalmann. Generic Personality and Emotion Simulation for Conversational Agents. Computer Animation and Virtual Worlds. 15(1): pp. 1-13, January 2004

H. Kim, T. Di Giacomo, A. Egges, E. Lyard, S. Garchery, N. Magnenat-Thalmann, " Believable Virtual Environment: Sensory and Perceptual Believability",Believability in Virtual Environment, December 2004

H. Kim, C. Joslin, T. Di Giacomo, S. Garchery, N. Magnenat-Thalmann, Adaptation Mechanism for Three Dimensional Content within the MPEG-21 Framework, Computer Graphics International 2004, June 2004

G.Papagiannakis, S. Schertenleib, B. O’Kennedy , M. Poizat, N.Magnenat-Thalmann, A. Stoddart, D.Thalmann, “Mixing Virtual and Real scenes in the site of ancient Pompeii”, Journal of CAVW (to appear), November 2004.

I-1: Introduction to Mixed Reality

Daniel Thalmann

Virtual Reality Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)

Abstract

We will first explain what is Mixed Reality covering the spectrum from Reality to Virtual Reality. We will emphasize Augmented Reality which augments the user's view of the real world by composing 3D virtual objects with their real world counterparts, necessitating that the user maintains a sense of presence in that world. After a State-of-the-Art of the main technologies and concepts, we will survey a few applications in medicine, maintenance, tourism, and cultural heritage

.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Three-Dimensional Graphics and Realism]: Augmented Reality, Virtual Humans, Mixed Reality

1. Mixed Reality from Real to Virtual

We know that Virtual Reality immerse a user inside a virtual world that completely replaces the real world outside. Augmented Reality augments the user's view of the real world by composing 3D virtual objects with their real world counterparts, necessitating that the user maintains a sense of presence in that world. In fact, Mixed Reality (MR) corresponds to a complete spectrum (see Figure 1) from the Real Environment to the Virtual Environment through Augmented Reality and Augmented Virtuality where the World is Virtual with a few pieces of Reality.

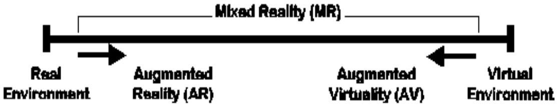


Figure 1. From Real to Virtual

Augmented Reality is currently an active research topic as well as a high potential commercial application target. As researchers continue to improve the tracking, display and mobile processing components of MR systems, the seamless integration of virtual and sensory information may become not merely possible but commonplace. Many observers have suggested that one of the many potential applications of augmented and mixed realities will emerge as the “killer app”- a use so compelling that it would result in mass adoption of the technology.

However, the topic of integrated virtual human simulation in augmented reality is not explicitly covered yet in the current bibliography or the latest patents list. Azuma [ABB*] describes an extensive bibliography on current state-of-the-art AR systems & frameworks. However, few of these systems take the modern approach that a realistic mixed reality application, rich in AR virtual character experiences, should be based on a complete VR Framework (featuring game-engine like components) with the addition of the “AR enabling Technologies” like a) Real-time Camera Tracking b) AR Displays and interfaces c) Registration and Calibration.

For creating scenes involving, for example, virtual actors in the real world, we should really take into account the real world during the generation of the images by the computer. For example, consider a virtual actor passing behind a real tree: for some images of the actor, part of the body should be hidden. For more realism, the shadow of the actor should be cast on the real floor. This means that the computer-generated images are dependent on the real world. One way of solving these problems is to create virtual objects similar to the real ones and a virtual camera corresponding to the real camera which is used to shoot the real scene. However, this correspondence is generally hard to establish.

In summary, the virtual actor should be integrated into the real world using the same parametric conditions than in the reality. This means that several interesting problems should be solved [MT97]:

- “collision detection” between the virtual actor and the real environment; e.g. virtual actor walking on a real street or sitting down on a real chair.
- processing the hidden surfaces which means real objects hidden by virtual actors and virtual actors hidden by real objects.

- adapting the sizes of the virtual actors to the dimensions of the real world.
- making the rendering of the virtual actor similar to the representation of the real world (photo or video)
- casting shadows of the virtual actors on the real world
- if there is a camera motion, making a correspondence between the virtual camera and the real one (camera calibration)

2. Hardware for MR

2.1 Optical See-Through HMD

One way to implement Augmented Reality is with an optical see-through Head-Mounted Display. This device places optical combiners in front of the user's eyes. The combiners let light in from the real world, and they also reflect light from monitors displaying graphic images. The result is a combination of the real world and a virtual world drawn by the monitor

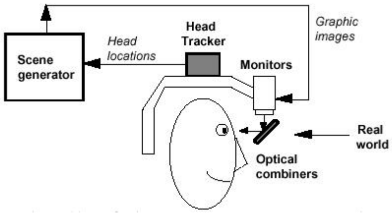


Figure 2. Principle of optical See-Through HMD

2.2 Video See-through Augmented Reality Display

The user's view of the real world is provided for the video cameras. The scene generator creates graphic images that are combined with the video, merging the real and virtual. The result is sent to the monitors. The video composition can be done through chroma-keying or depth processing.

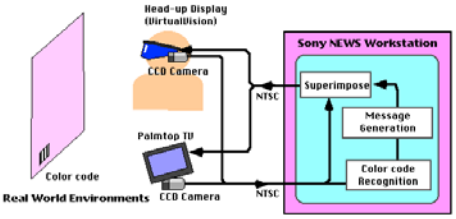


Figure 3. The Sony NAVICAM Video See-through Augmented Reality Display

We can consider, for example, the SONY Navicam system shown in Figure 3.

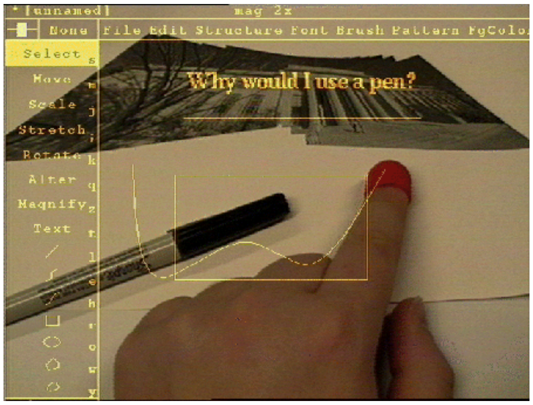
3. Applications of Augmented Reality

A growing number of projects are currently based on AR integrated platforms, exploring a variety of applications in different domains such as medical [ART04], cultural heritage [SDS*01 [GSO05], training and maintenance [SFCV01] [WT00] and games [TCD*00]. Special focus has recently been applied to system design and architecture in order to provide the various AR enabling technologies a framework [GHJ94] for proper collaboration and interplay.

We just now present a few examples.

3.1 Finger Tracking

Finger Tracking is one of the simplest application of Augmented Reality, the computer can visually track the user's finger (see Figure 4), there is no need to use a pen, a mouse or other devices



Why would I use a pen?

Figure 4. Finger Tracking

3.2 Annotation and visualization

AR could be used to annotate objects and environments with public or private information to aid general visualization tasks. One might be able to look out a window and see how a proposed new building would change her view.

3.3 Augmented Museum

The system detects the ID of a picture, and generates a description of it. For example, NaviCam identifies which picture the user is looking at and displays relevant information on the screen (Figure 5).

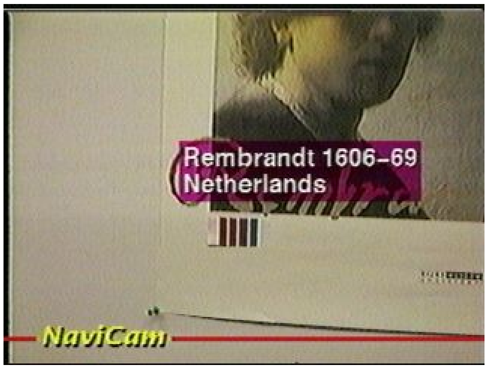


Figure 5. Augmented Museum

3.4 Medical

This domain is viewed as one of the most important for AR Systems.

AR could aid the doctors in the visualization and training for surgery.

AR may provide an internal view of the real patient.

- Through non-invasive sensors like Magnetic Resonance Imaging (MRI), Computed Tomography scans (CT), or ultrasound imaging could collect 3D datasets of a patient in real time.
- These datasets are rendered and combined in real time with a view of the patient, giving a "X-Ray vision" of the patient for doctor.

3.5 Manufacturing and repair

AR technology could provide instructions that might be easier for user to understand an equipment.

Theses instructions are not available in manuals with text and pictures, but as superimposed 3D drawings upon the actual equipment.

Theses drawings show step-by-step the tasks that need to be done and how to do them.

Instructions for assembly, maintenance and repair of complex equipment:

- Aircraft [Boeing]
- Printers [e.g. at Columbia University]
- Engines
- Automobile assembly

3.6 Mobile Augmented Reality Systems

It explores the synergy of two promising fields of user interface research:

- Augmented reality (AR), in which 3D displays are used to overlay a synthesized world on top of the real world, and
- mobile computing, in which increasingly small and inexpensive computing devices and wireless networking allow users to have access to computing facilities while roaming the real world

3.7 The Touring Machine

This system (Figure 6) developed at Columbia University compiles in a single system a HMD, a tracking device, and a mobile CPU (Central Processing Unit) - created at Columbia University.

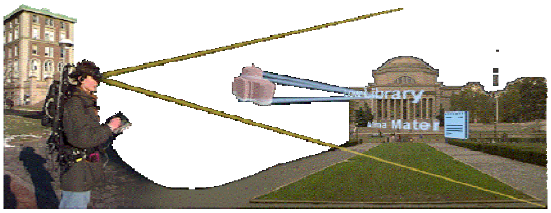


Figure 6. The Touring Machine (Columbia University)

3.8 Augmented Simulation (AUGSIM)

AUGSIM combines Augmented Reality (AR) and Seamless Simulation to augment conventional live and simulator based exercises for training and gaming.

Live exercises can be augmented with computer simulated entities and actions. Simulator exercises can be augmented with live entities and actions. Two-way real-time Seamless Simulation allows live and virtual entities to exist and interact in the same exercise Augmented Reality allows virtual entities and actions to be seen and heard integrated into the real world. Thus, AUGSIM allows live entities to see and hear virtual entities and actions, and virtual entities to see and hear live entities and actions.

3.9 Virtual Heritage

This is one of the most promising areas as it may provide Augmented Life to old historical sites, as for example Roman life in Pompei, as shown in Figure 7 from the LifePlus project [GSO05].

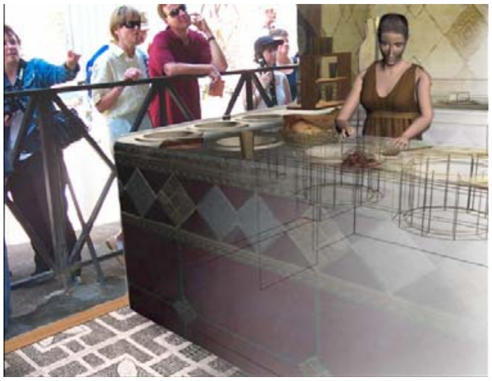


Figure 7 Augmented Reality in Pompei (LifePlus project)

References

[MT97] Magnenat Thalmann N., Thalmann D., Animating Virtual Actors in Real Environments, ACM Multimedia Systems, Springer,Vol.5, No2, 1997, pp.113-125.

[ART04]. ART: Augmented Reality for Therapy, http://mrcas.mpe.ntu.edu.sg/groups/art/, last accessed: 27/08/04

[SDS*01] Stricker, D., Dähne, P., Seibert, F., Christou, I., Almeida, L., Ioannidis, N., Design and Development Issues for ARCHEOGUIDE: An Augmented Reality-based Cultural Heritage On-site Guide, *EuroImage ICAV 3D Conference in Augmented Virtual Environments and Three-dimensional Imaging*, Mykonos, Greece, 30 May-01 June 2001

[GSO05] Papagiannakis G., Schertenleib S., O'Kennedy B., Poizat M., Magnenat-Thalmann N., Stoddart A., Thalmann D., Mixing Virtual and Real scenes in the site of ancient Pompeii, *Computer Animation and Virtual Worlds*, Vol.16, No1, 2005 pp.11-24

[SFCV01] Schwald, B., Figue, J., Chauvineau, E., Vu-Hong, F., STARMATE:Using Augmented Reality technology for computer guided maintenance of complex mechanical elements, *e2001 Conference*, 17-19 October 2001 - Venice – Italy

[WT00] Wohlgemuth, W., Triebfürst, G., ARVIKA: augmented reality for development, production and service, *Proceedings of DARE 2000 on Designing augmented reality environments*, 2000, Elsinore, Denmark

[TCD*00] Thomas, B., Close, B., Donoghue, J., Squires, J., De Bondi, P., Morris, M., and Piekarski, W. ARQuake: An Outdoor/Indoor Augmented Reality First Person Application. *In 4th Int'l Symposium on Wearable Computers*, pp 139- 146, Atlanta, Ga, Oct 2000

[GHJ94] Gamma, E., Helm, R., Johnson, R., Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1994

[ABB*] Azuma, Baillot, Behringer, Julier, MacIntyre, Recent Advances in Augmented Reality, *IEEE Computer Graphics and Applications*, Nov/Dec 2001

I-2: Believability and Presence

Nadia Magnenat-Thalmann, HyungSeok Kim, Georgios Papagiannakis, Thomas Di Giacomo

MIRALab, University of Geneva, Geneva, Switzerland

Abstract

In the Mixed Reality (MR) environment, the concept of cyber-real space interplay invokes such inter-affective experiences that promote new patterns of believability and presence. Believability is a term used to measure the realism of interaction in the MR environments. Presence is defined as the measure that is used to convey the feeling of 'being there'. In this session, a concept of Believability and Presence for MR is presented, starting from the sensory level interaction to the perceptual level interaction, focused on the inhabited environment. We show an example of inhabited MR environment which show strengthened presence but without enough believability, due to limited interaction between the real participants and the virtual characters. Among present technologies for MR environment, we introduce a new concept of 'affective registration' to enhance sensory level experience through keeping consistency of the virtual environment with the real environment. We also argue that future steps in MR enabling technologies should cater for enhanced social awareness of the virtual humans to the real world and new channels for interactivity between the real users and virtual actors. Only then the believability of interactions in a MR environment will be enhanced and allow for compelling real experiences.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Three-Dimensional Graphics and Realism]: Believability, Virtual Reality, Virtual Humans, Mixed Reality, Presence, Emotion

1. Introduction

The interaction process in the MR environment can be described as a process to exchange information between the real participants and the virtual world. The virtual world includes mixed-reality environment along with virtual humans and other participants. In terms of traditional 'realism', that information mainly consists of sensory information such as visual information, auditory information, and haptic information. Focused researches on those issues enable an enhanced level of realism in many application areas, but it is also found that this conventional concept of realism does not guarantee the 'realistic' interactions. For example, very 'unrealistic' representation on a game can give high level of 'realism' while very 'realistic' representation for the virtual world navigation application attracts little interest from participants. In this session, we present concept of believability along with presence to cover this gap.

Believability is a term to measure a level of realism in experiences in the interactive MR environment. Presence is defined as the measure that is used to convey the feeling of 'being there'. Believability is a different but dependent concept with presence. An experience of realistic interaction could give high level of presence and a high level of presence could trigger a believable interaction. In this session, main focus will be given into a concept of believability, and the presence is viewed as an element of this concept. Other aspect of presence which is mainly focused on presentation of sensory information is not covered in this session.

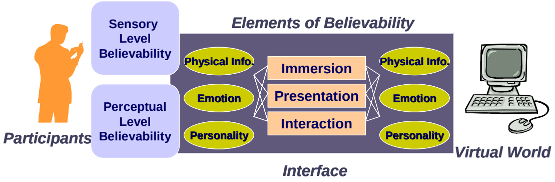


Figure 1 Concept of the believability

The definition of believability is still an open issue. Zeltzer states that Autonomy, Interactivity, and Presence are important elements for the Virtual Environment [Zel92]. These elements are one of the most essential ones to make the virtual world `realistic' but in terms of believability, a traditional definition of these terms is not sufficient. For the character representation, the believability is often discussed in context of generating behaviors [LS97]. Believable behavior covers not only realism but also emotions, personality, and intent [Doy02]. The believable experience can be defined as a set of interactions which gives perception of properties of the environment [KdGE*04]. Figure 1 is an exemplar illustration of the believable interaction. Perceptual properties, such as physical information and emotional information, are exchanged among participants and the virtual worlds by the interpretation of the interface. Those interpretations can be categorized as immersion, presentation and interaction.

- Immersion

The user can believe that the experience in the virtual world is a real experience if he or she is totally immersed in the virtual environment. Modeling and measuring the immersion has been conducted by utilizing both cognitive surveys and performance indicators [PPW97][RCvZ97]. In other points of view, users are immersed into the virtual world if their experiences are realistic. Sheridan called this element as `the active imagination in suppressing disbelief (and thus enhanced believability)' [She00]. The semantic experience in a virtual environment consists of emotional elements, personalized elements and goal-oriented elements. If a set of objects and stories have these elements, participants believe the represented world. This level of immersion is often called *presence*.

- Presentation

The believability of the virtual environment can be increased if the virtual world is presented as real as real world. The realism in the interactive process is not only deal with level of similarity to the real world for each sensory channel, but also concerned in synchronized level of realism between multisensory channels. For example, in some cases, extremely high level of realism on one sensory channel may hurt the entire experience if it is accompanied with low level of realism in other sensory channels.

- Interaction

One of the most important issues in the virtual environment is its interactivity. A realistic interactive system will result in higher believability in normal cases. The sensory feedback should be fast enough from its corresponding action input. Also it should be fast enough to the human visual sensor. In computer gaming environment, it is well known fact that slow visual refresh rate will hurt performance compared with fast refresh rate. In addition to the fast interactivity, the realism of the interactivity can be determined by its behavior. The interactivity is increased if the behavior responds to actions of users in a life-like way.

Realistic reactive behavior in interactivity is related but different from behaviors to induce perceptual immersion. Immersion largely depends on how well this is implemented, for example through goal-oriented artificial intelligence or emotional behavior simulation [MP99][LS97]. We believe that the perceptual immersion is invoked by goal-oriented intervention of intents, emotions, and personality. The realism of the interaction is defined by the involvement of the user in the virtual environment. For example, factors of presence as defined by Stevens et al. [SJH02] are re-categorized so that: 1) personal presence, intended aspect of social presence and task factors are components of immersion, 2) unintended aspect of social presence and environmental presence are components of realism in interactivity.

These effects of these elements are not independent. They influence each other in a complex way. In some cases a high level of realism for one area will elaborate the level of believability but if it is combined with a low level of realism on other area, it will decrease the level of believability. Even if the sensory channel has enough realism, it is not sufficient to make the VE believable if the VE does not have believable contents. From another point of view, a VE presented in written text (for example a novel or a book) depending on the quality of the stories.

Perceptual properties of the virtual environment are further layered into objective-perceptual and subjective-perceptual ones. A set of experiments and its results are illustrated to identify a condition to trigger a believability to perceive those parameters in multi-modal environment. Throughout experiments perception level of objects is analyzed with different modalities, different realism for the properties and manipulated/missing properties.

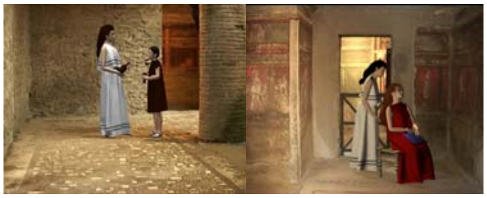


Figure 2 Inhabited MR environment example [Lifeplus project]

The last but not the least important aspect of interaction is its consistency and persistency. To make the interaction believable, the process should be within the ‘expectation’ of the participant. We define this constraint as *consistency of the interaction* with respect to the participant’s expectation. This is different with the constant reaction. For example, if the virtual character in the MR environment (Figure 2) only shows the same greetings to your greeting behavior, it does not give enough believable experience. The virtual character should recognize you and react to you according to their emotional and personalized status, your status and environmental aspect.

In addition, the believability can be triggered by ‘persistent’ interaction. The result of action made by participant should last for the next occasion. The meaning of ‘last’ does not means preservation of the status, but means expectable change on the status due to interaction made by the participant, other participants and the environment.

2. Sensory level issues to enhance presence and believability

In terms of interface, believability can be discussed for each sensory channel. Among primary human sensory channels visual, auditory and haptic has been major elements in terms of interface.

In the MR environment, in addition to the previously mentioned aspect, one additional constraint is given as the interface should not interfere with the real-environment.

With well provided sensory information, it is not only possible to trigger believability by transferring perceptual information from the MR environment to the real participants, but also provide 'illusions' to the user. Annie Luciani et al showed that if suitable physical information is given, illusion on the physical status can be triggered even with non-realistic representations [LUM04].

In this section, the believability issues of major three major channels are discussed.

2.1. Realistic visual sensory feedback

The visual sensory channel is a one of the most important channel to make virtual world believable. For example from the early version of movies, it has given successful believable experiences to audiences using mostly visual information only. Visual channel is the most investigated sensory channel in the virtual reality scene. Issues including modeling and re-producing visual information are investigated since the beginning of the computer graphics in 60's. They are started from the modeling and re-producing the virtual world itself and it is evolved to integrating real and virtual world altogether.

Visual immersion is achieved through the use of see-through or video-through displays in the MR environment. The high level of immersion requires accurate registration and real-time feedback. There have been many works to measure 'sense of presence' for difference visual immersion levels. These are measured in terms of distance/depth perception, task performance, and easy of use.

To achieve realism in the presentation, most of work has been done to generate images to have image level realism. The image level realism is defined as a state of realism in image with comparison in the real-image in terms of pixel-wise comparison. Realistic shape modeling and realistic illumination modeling fall into this category. In the MR environment, important aspect is sensory level consistency. The sensory level information for synthesized objects should keep consistency with the sensory level information in the real environment.

Recently, there have been some approaches to consider human sensory limitation or perceptual issues such as give more detailed model where human visual sensor can perceive its delicate details.

2.2. Reproducing auditory information for MR environment

The audio is as or even more important than the video. The surrounding sound defines the environment all around the participant. Again, the problem of the auditory information generation is both on complexity and fidelity in modeling and rendering process.

3D spatial audio in MR environments is a relatively new and wide research topic, although spatial audio in general has been under investigation since the beginning of the last century. Rendering audible space with preserved three-dimensional sound illusion is called auralization according to Kleiner[KDS93]. Virtual acoustics include virtual reality aspects such as dynamic listener, dynamic source and acoustic environment specificities [THV*96] [FMC99] [SHLV99]. Some fundamental elements already existent are necessary for a complete spatial audio system including transmission, reflections, reverberation, diffraction, refraction and head related transfer. As can be observed, some of these elements are affected by the position of the sound source relative to the listener (or receiver) and others are affected by the environments itself. Several propagation methods are proposed to simulated sound effect from the sound source to the listener [Bor84][Kul84][FTC*04]. Most of sound rendering techniques reproduces the sound field for a specific listening point.

In the MR environment, the consistency and persistency aspect of believable experience comes to the consistent effects of environment to the reproduced sound. Often, the sound environment can be modeled as very simplified 3D environment extracted from the real-environment. Using this approach, it is possible to generate a sound inside the MR environment.

2.3. Haptic interaction in MR environment

Until now, haptic sensory feedback is simulated in limited way especially for the MR environment. Although there have been discussion and illustration on full body

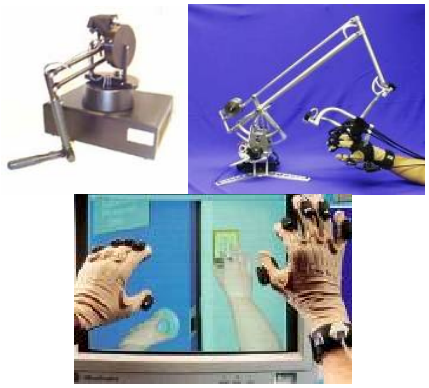


Figure 3. Examples of haptic devices. From the top-left, arm-like device [Pha], exoskeleton [Imm], and tactile display [Imm]

haptic reproduction, for example data suite, the current technological level is still far away from that goal. Currently, most effort is devoted to simulate realistic presentation of haptic. So far, common haptic devices are not suitable for the MR experience due to its limit on the portability and interference with the real-environment.

2.4 Multisensory issues

A virtual environment is an interactive system in which the user manipulates and experiences a synthetic environment through multiple sensory channels. In a multimodal system communication channels are numerous: voice, gesture, gaze, visual, auditory, haptic etc. Integrating these modalities (multimodal inpu

ts) improves the sense of presence and realism and enhances human computer interaction. Virtual environments using sight, sound and touch are quite feasible, but effects of sensory interaction are complex and vary from person to person. Nevertheless adding several communication channels leads to system complexity, cost, and of integration/synchronization problems. Sensory Fusion is a relatively new topic, for which we need to study two kind of human computer communication channels.

In addition to the issues in the uni-sensory channel, the multimodal interface introduces more complex situation to invoke believability of the environment

Among issues of the believability of virtual environment, one of the key issues in the multimodal interface is synchronization of sensory channel. Some anomalies such as motion sickness are appeared when synchronization of sensory channel is not well provided. The basic question on this is the tendency of the believability to the synchronization level.

The rate of sensory feedback is another issue that affects believability of the interface. It is known that the visual sensory channel should provide feedback at a speed of around 60 frames per second to give enough realism. But when it combined with action channel sometimes speed of 100 frames per second is not enough in some cases. The question should be answered is identifying effective elements of sensory channel that affects to the other channel.

The most difficult but interesting issue is the difference in the level of realism among sensory channels. Even if one sensory channel could provide very realistic informational feedback, other channels with low realism could prevent to create high level of believability. In other cases, there have been a set of research that one sensory channel could replace other sensory channel. It may suggest that there might be discrimination on the differences of the realism between sensory channels. The question on this issue is discriminating the possible threshold where the higher realism in a specific sensory channel is desired.

3. Animation Believability with Motion Perception

In this section, we discuss psycho-physical approaches of motion perception by the human visual system, and its potential impact on believability for Computer Graphics and Animation. After basic neurophysiologic mechanisms of perception and low-level motion detection and discrimination, higher-level concepts such as motion memory and learning, observer attention, are explored in the context of believability for Graphics.

3.1 Motion Detection

Three complementary approaches coexist in perception [SWB02]: computational, psychophysical and neurophysiologic. Physiologically, the detection and analysis of motions are processed through a cascade of neural operations located in different areas of the brain: after registering local motions, interconnected neuron areas communicate to merge local signals into global descriptions of motion. Important terms from the motion perception literature are:

Aperture: Similarly to receptive fields, it is an opening within which neurons register motion. It can be seen as a spatially restricted window or a viewpoint in graphics.

Stimulus: From [Ste51], "The complete definition of the stimulus to a given response involves the specification of all the transformations of the environment that leave the response invariant". When ensuring these invariant conditions, stimuli such as Gabor patterns (sinusoidal gratings and Gaussian functions), moving lines or dots, or computer animations, allow the understanding of response processing.

Optic Flow: It is a continuous sequence of images perceived by humans due to spatio-temporal changes. It provides information on shapes, distances, velocities, *etc.* and drives path-finding with collision awareness and avoidance, as well as depth segregation.

Additional results are provided in [Ade91] and in other work. The lower limit for detection, *i.e.* motion acuity, varies with the number of objects. The detection of relative and biological motions, such as human gait, is far better than absolute or random motions. Motion detection varies with the size, exposure duration of the stimuli, but not with the direction. Exposures to coherent motion, *i.e.* motion with a general direction, also increases the motion discrimination, *i.e.* distinction between trajectories. The constant properties of velocity perception has been studied in [DGvVH00], the smoothing of motion perception in [RES01], and it has been shown that motions based on translation are predictable by human brains. Most of the time, objects are subjected to different combinations of transformations, series of translations and rotations. [YSP02] study the perception of objects rotating and translating, with lines as stimuli. They conclude that motion perception of rotation and translation is largely independent of the aperture shape configuration.

3.2 High-Level Perception and Believability

Memory for motion is a relatively new topic of research and though work has illustrated robustness for memory of velocity over 30 seconds or coherence in memory for direction, many questions still remains. For instance, though high level features have an impact on motion perception, this impact is minor and under restricted conditions of information semantics: [Yu00] shows that semantic knowledge of objects is influencing motion correspondence. Other senses and information are as well modifying visual and motion perception, such as sound for instance, as detailed in [SSL97]. Related to attention and interest, preferences for scenes are explored in [VB01]. Based on the assumption that we do not look at random spots, their experiments explore our internal mechanisms for the selection of regions of interest. Research has been conducted on recognition of human motion such as arm lifting or gait. For instance, the genre and the individuality are potentially recognized with gait as stated in [Tro02]. Though it is highly articulated and deformable, the human body movements are very easily discriminated by the visual system of human observers. Following experiments on the integration of human multiple views in motion, [KS99] suggest that human movement perception might be based on the body biomechanical limitations, and potentially confirm that the perception of motion and the object-recognition process are closely linked. [HOT98] have explored the recognition of human motion, with computer animated sequences as stimuli. Applied perception is often used as a criterion for level of details methods and therefore to impact on the representation believability. Early work of [FS93] proposes adaptive algorithms for visualization at stable frame rates, according to the size, focus, motion blur and semantics of perceived objects. [OD01] use perception for the believability of real-time collision detection. For improving believability, perception is also a mechanism to validate the level of complexity of physically-based animation, as explored in [BHW96] and later in [ODGK03]. Generally, [Mys02] detail some perception-based metrics for walkthroughs in credible virtual environments while [RP03] propose such an approach for believability of virtual human animation by the detection of non-plausible and plausible motions.

4. Affective registration: an example of illumination registration

Attaining a high quality believability and realism of a real-time seamless integration between real images and virtual objects lit with ‘real light’, requires two main aspects for consistent matching: geometry and illumination. First, the camera position-orientation and projection should be consistent; otherwise the object may seem too foreshortened or skewed relative to the rest of the picture (geometrical consistency). Secondly, the lighting-shading of the virtual object needs to be consistent with the other objects in the real environment (illumination consistency). In the past, consistency of geometry has been intensively investigated [ABB*01]. On the other hand, few methods have been proposed so far for real-time illumination consistency registration to superimpose virtual objects onto an image of a real scene: [JL04], [FGR93] and even less research in the area of superimposing real-time, dynamic-deformable virtual objects on real-time video streams. Furthermore, pursuing bridging the lighting of virtual objects with real ones, we harness the research graphics domain of global illumination, the physically-correct simulation of light transport [WPS*03], [DDM02].

With the advent of sixth generation graphic programmable units (GPUs) as massive parallel powerful streaming processors, research on real-time ray tracing has recently made tremendous advances [WPS*03]. Recent algorithmic improvements together with optimized GPU based implementations allow now for limited interactive ray tracing. Furthermore, as most of today’s global illumination algorithms heavily build on top of ray tracing, real-time performance of the latter is giving rise to new interactive global illumination algorithms for complex dynamic scenes [DDM02]. However, the application of such models is still far away from MR and dynamic virtual cultural heritage, due to their heavy computational requirements. Recently a new real-time theory-methodology for physically correct area-light global illumination simulation of rigid objects in VR has been introduced, termed Precomputed Radiance Transfer(PRT) [SKS02], which provided the most realistic and believable real-time VR illumination model up to date [Kau04]]. In the current work we propose an extension on this algorithm for a) MR (matching exposures of real light captured from real-time AR camera and real-light captured from light probe) and b) multi-geometry meshes and a categorization of the radiance transfer based on the type of sub-geometry in the mesh: occluder or receiver (Figure 4). In [PLF*01] a radiosity based solution with irradiance maps was presented for photorealistic virtual heritage static object simulation, but allowed only static, predefined objects, lights where the currently introduced VR PRT algorithm allows for real, dynamic High Dynamic Range Image (HDRI) lights in a physically correct simulated environment.

The new proposed fusion of developments in this area of real-time physically-correct simulation of light transport (low frequency shadows, reflections, indirect illumination) with the previous topic of illumination registration in Augmented Reality environments, we believe will become a mandatory feature of future Mixed Reality simulations, similar to the introduction of real-time texture mapping a few years ago.

Our belief is that Mobile MR can be a better vision for the future if the above shortcomings are met so that both notions of believability as well as presence can be reinforced.

So far, previous approaches regarded believability as related more closely with the platonic notion of inversed world of senses – ideas respectively used to represent the virtual-real world. In that representation, believable is what imitates reality (ideal) whereas actual MR experience is paralleled to the flawed sensual world.

In this work we have addressed the issues of creating interactive applications for mobile MR, in order to deliver 'real' experiences where illumination registration is addressed with captured 'real-light'. We believe that further synergies between Knowledge Media, Semiotics of Presence and Hermeneutical Phenomenology will help to establish a theoretical framework of the 'signs' of Believability and Presence in MR media. Furthermore, recent state-of-the-art research in the areas of neuroscience and psychological models can provide the needed clinical and physiological evidence. Only then MR, Media, Vision and Wearable computer scientists will be able to capitalize on the foundations of Believability and Presence for extending the virtuality MR structures and enabling compelling real experiences through mobile virtual environments.

5. Perceptual level Presence and Believability: Introducing the interactive future

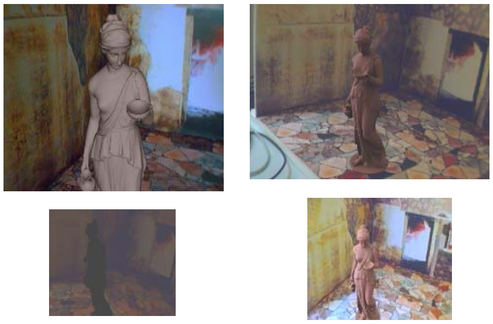


Figure 4. Virtual augmentation in AR: Normal diffuse shading (top left) and extended Precomputed Radiance Transfer with varying exposure of both h/w camera and captured HDRI scene real area light (bottom left, right)

As human beings are from experience more used to interaction with other human beings, creating interfaces that mimic human-like behavior has become a very important research topic. Such kind of behavior depends heavily on the quality of perception, the underlying interactive techniques and the expressive capabilities of the interface. In the MR environment, we are now able to create realistic virtual characters that can move and talk. Multi-modal interaction with such characters is possible as the required technologies are getting mature (speech recognition, natural language dialogues, speech synthesis, animation, and so on). All these different technologies that are available to us nowadays, determine in a major way the *believability* of the human-like interface. The believability can be measured by looking at the expressive functions of the interface (does the human-like interface look like a human?), the internal functions (does the human-like interface portray human-like behaviour?), and the perceptive functions (does the human-like interface perceive like a human does?). On the other hand, one can question that a believable interface necessarily has to be human-like. Some cartoon-like interfaces can be very engaging, while they do certainly not look like humans or behave like them. A more general way of looking at believability is by defining it as a means to evaluate how well the different parts of the interface integrate and if the interface is controlled in a meaningful and consistent way. This 'driving force' behind the interface can be compared with the *individuality* of human beings. Two important factors that help to define this individuality are personality and emotion. These factors act as glue between perception, dialogue and expression.

There are different scenarios that describe how an emotion is evoked from the perception of one or more events (see Figure 5 for an overview). The process of inducing an emotional response from perceptive data is called **appraisal**. One of the oldest theories, the James-Lange theory of emotion states that an event causes arousal first and only after our interpretation of the arousal, we experience an emotion. The Cannon-Bard theory of emotion [Can27] states that emotion and the physiological response happen at the same time and unrelated from each-other. The Schachter-Singer scenario [SS62] says that an event causes arousal, but that the emotion follows from the identification of a reason for the arousal. The Lazarus theory of cognitive emotion [Laz91] states that both arousal and emotion are invoked separately by a thought following an event. Finally, the Facial Feedback hypothesis [Buc80][LCK76] says that emotion is the experience of changes in the facial muscle configuration. This result has also been shown by Ekman et al. [ELF83].

In emotion simulation research so far, appraisal (obtaining emotional information from perceptive data) is popularly done by a system based on the OCC model [OCC88]. This model specifies how events, agents and objects from the universe are appraised according to respectively their desirability, praiseworthiness and appealingness. The latter three factors are decided upon by a set of parameters: the goals, standards and attitudes. The model delivers us emotional information (i.e. the influence on the emotional state) with respect to the universe and the things that happen/exist in it. In order to have a working model for simulation, one is of course obliged to define the goals, standards and attitudes of the simulator. These factors can be considered as the ‘personality’ of the simulator. In this case, the personality of a simulator is (partly) domain-dependent. However, more recent research indicates that personality can be modeled in a more abstract, domain-independent way [Eys90] [CM92]. In this case, personality is a set of factors/dimensions that each denote an influence on how perception takes place and how behavior is shown.

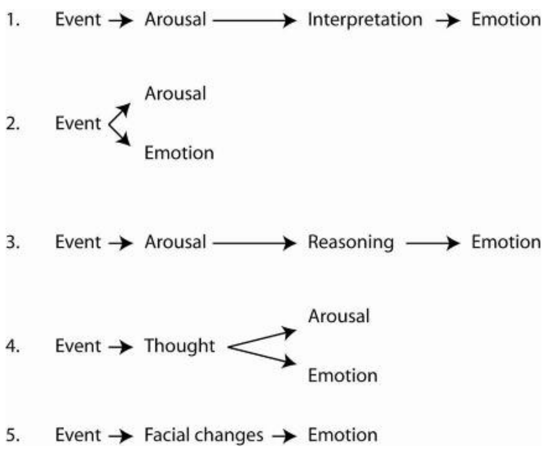

1. Event → Arousal → Interpretation → Emotion
2. Event
Arousal
Emotion
3. Event → Arousal → Reasoning → Emotion
4. Event → Thought
Arousal
Emotion
5. Event → Facial changes → Emotion

Figure 5. Five scenarios to describe the path from event to emotion: (1) James-Lange (2) Cannon-Bard (3) Schachter-Singer (4) Lazarus (5) Facial Feedback.

Egges et al. [EKM04] shows a system that tries to integrate multi-dimensional, domain-independent personality models and appraisal models.

The effect of personality and emotion on agent behavior has been researched quite a lot [Piw02], whether it concerns a general influence on behavior [MG02], or a more traditional planning-based method [JS01]. Various rule-based models [AKG*99], probabilistic models [BB98] and fuzzy logic systems [EIY99] have also been developed.

Finally, personality and emotion will have an effect on how behavior is expressed (speech will have different intonations, a face will make expressions reflecting the emotional state, a body will make different gestures according to the personality and the emotions). A first step towards such an approach is described in [KM02], which describes a system that can simulate personalized facial animation with speech and expressions, modulated through mood. There have been very few researchers who have tried to simulate mood. Velasquez [Vel97] proposed a model of emotions, mood and temperament that provides a flexible way of controlling the behavior of the autonomous entities. According to Velasquez, moods and emotions are only differentiated in terms of levels of arousal. He proposed simple models for blending, mixing and decaying emotions to subsequently select actions of the agent. Moffat [Mof95] states that personality and emotion are basically the same mechanisms only differentiated by two cognitive variables time and duration, and personality can be seen as consistent expression of emotion.

The influences of personality and emotion on perception, interaction and expression of human-like interfaces will certainly have an effect on the believability of the interface. Depending on the different models, in which combination they are used, and depending on how the link between personality/emotions and the interface is established, the believability of the interface will be affected. Therefore it is crucial to define an efficient, objective way to evaluate such interfaces depending on so many different parameters. One needs to define metrics that are independent of the approach that is used to create the interface and that are still capable of telling us whether an interface is believable or not.

6. Conclusion

The believability is a measure of essential realism required for the interaction process. In the MR environment, the believable interaction could be achieved through consistent and persistent transfer of perceptual information. The consistency mainly covers issues of the reaction of MR environment to the actions and expectation of the participants. In sensory channel level, the consistency also includes consistency of the real and virtual sensory feedback especially for the MR environment.

The persistent interaction can be enhanced by introducing virtual humans in the MR environment. The virtual characters which are recognizing emotion and personalities of real participants, and which behaves emotionally consistent and persistent way, will enhance the whole experience in believable way.

Acknowledgement

Part of the work presented has been supported by the EU IST FP6 programme, in frame of the EU IST ENACTIVE (http://www.enactivenetwork.org/) project.

References

[Zel92] ZELTZER D.: Autonomy, interaction, and presence. Presence: Teleoperators and Virtual Environment, 1(1):127–132, 1992.

[LS97] LESTER J. C., STONE B. A.: Increasing believability in animated pedagogical agents.

In Proceedings of the first international conference on Autonomous agents, pages 16–21. ACM Press, 1997.

[Doy02] DOYLE P.: Believability through context using 'knowledge in the world' to create intelligent characters. In Proceedings of the first international joint conference on Autonomous agents and multiagent systems, pages 342–349. ACM Press, 2002.

[KdGE*04] KIM H., DI GIACOMO T., EGGES A., LYARD E., GARCHERY S., MAGNENAT-THALMANN N.: Believable Virtual Environment: Sensory and Perceptual Believability, International Workshop on Believability in Virtual Environment, December 2004

[PPW97] PAUSCH R., PROFFITT D., WILLIAMS G.: Quantifying immersion in virtual reality. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pages 13–18. ACM Press/Addison-Wesley Publishing Co., 1997.

[RCvZ97] ROBERTSON G., CZERWINSKI M., van Dantzich M.: Immersion in desktop virtual reality. In Proceedings of the 10th annual ACM symposium on User interface software and technology, pages 11–19. ACM Press, 1997.

[She00] SHERIDAN T. B.: Interaction, imagination and immersion some research needs. In Proceedings of the ACM symposium on Virtual reality software and technology, pages 1–7. ACM Press, 2000.

[MP99] MARTINHO C., PAIVA A.: Pathematic agents: rapid development of believable emotional agents in intelligent virtual environments. In Proceedings of the third annual conference on Autonomous Agents, pages 1–8. ACM Press, 1999.

[SJHC02] STEVENS B., JERRAMS-SMITH J., HEATHCOTE D., CALLEAR D.: Putting the virtual into reality: assessing object-presence with projection-augmented models. Presence: Teleoper. Virtual Environ., 11(1):79–92, 2002.

[LUMC04] LUCIANI A., URMA D., MARLIERE S., CHEVRIER J.: PRESENCE : The sense of believability of inaccessible worlds. Computers & Graphics. Elsevier Ed.. Vol 28/4 pp 509-517, 2004.

[KDS93] KLEINER M., DALENBCK B.-I., SYENSSON P.: Auralization - an overview. Journal of the Audio Engineering Society, 41:861–875, 1993.

[THV*96] TAKALA T., VAANANEN R., VLIMKI V., SAVIOJA L., HUOPANIEMI J., HUOTILAINEN T., KARJALAINEN M.: An integrated system for virtual audio reality. In 100th Convention of the Audio Engineering Society, 1996.

[SHLV99] SAVIOJA L., HUOPANIEMI J., LOKKI T., VAANANEN R.: Creating interactive virtual acoustic environments. Journal of the Audio Engineering Society, 47(9):675–705, 1999.

[FMC99] FUNKHOUSER T., MIN P., CARLBOM I.: Real-time acoustics modeling for distributed virtual environments. In SIGGRAPH 1999 Conference Proceedings, 1999.

[Bor84] BORISH J.: Extension of the image model to arbitrary polyhedra. Journal of Acoustics of America, 75:1827–1836, 1984.

[Kul84] KULOWSKI A.: Algorithmic representation of the ray tracing technique. Applied. Acoustics, 18:449–469, 1984.

[FTC*04] FUNKHOUSER T., TSINGOS N., CARLBOM I., ELKO G., SONDHI M., WEST J., PINGALI G., MIN P., NGAN A.: A beam tracing method for interactive architectural acoustics. Journal of the Acoustical Society of America, 2004.

[Pha] The Phantom webpage: http://www.sensable.com/products/phantom_ghost/phantom.asp

[Imm] The CyberForce and CyberTouch from the Immersion: http://www.immersion.com/

[SWB02] SEKULER R., WATAMANIUK S., BLAKE R.: Perception of Visual Motion. In Stevens' Handbook of Experimental Psychology, Sensation and perception, Wiley & Sons, 2002.

[Ste51] STEVENS S.: Mathematics, Measurement and Psychophysics. In Handbook of Experimental Psychology, Wiley & Sons, 1951.

[Ade91] ADELSON E.: Mechanisms for Motion Perception. Journal of Optics and Photonics News, 2(8):24-30, 1991.

[DGvVH00] DISTLER H., GEGENFURTNER K., Van VEEN H., HAWKEN M.: Velocity constancy in a virtual reality environment. Perception, 29(12):1423-1435, 2000.

[RES01] RAO R., EAGLEMAN D., SEJNOWSKI T.: Optimal Smoothing in Visual Motion Perception. Neural Computation, 13(6):1243-1253, 2001.

[YSP02] YANG Z., SHIMPI A., PURVES D.: Perception of objects that are translating and rotating. Perception, 31(5):925-942, 2002.

[Yu00] YU K.: Can semantic knowledge influence motion correspondence? Perception, 29(6):693-707, 2000.

[SSL97] SEKULER R., SEKULER A., LAU R.: Sound alters visual motion perception. Nature, 385:308, 1997.

[VB01] VESSEL E., BIEDERMAN I.: Why do we prefer looking at some scenes rather than others? OPAM, 2001.

[Tro02] TROJE N.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of Vision, 2(5):371-387, 2002.

[KS99] KOURTZI Z., SHIFFRAR M.: Dynamic representations of human body movement. Perception, 28(1):49-62, 1999.

[HOT98] HODGINS J., O'BRIEN J., TUMBLIN J.: Perception of Human Motion with Different Geometric Models. IEEE Transactions on Visualization and Computer Graphics, 4(4), 1998.

[FS93] FUNKHOUSER T., SEQUIN C.: Adaptive Display Algorithm for Interactive Frame Rates During Visualization of Complex Virtual Environments. SIGGRAPH 1993.

[OD01] O'SULLIVAN C., DINGLIANA J.: Collisions and Perception. ACM Transactions on Graphics, 20(3), 2001.

[BHW96] BARZEL R., HUGHES J., WOOD D.: Plausible Motion Simulation for Computer Graphics. Eurographics Workshop on Computer Animation and Simulation, pp.183-197, 1996.

[ODGK03] O'SULLICAN C., DINGLIANA J., GIANG T., KAISER M.K.: Evaluating the Visual Fidelity of Physically Based Animations. ACM Transactions on Graphics, 22(3), 2003.

[Mys02] MYSZKOWSKI K.: Perception-based global illumination, rendering, and animation techniques. Spring Conference on Computer Graphics, pp.13-24, 2002.

[RP03] REITSMA P., POLLARD N.: Perceptual Metrics for Character Animation: Sensitivity to Errors in Ballistic Motion. SIGGRAPH 2003.

[ABB*01] AZUMA R., BAILLOT Y., BEHRINGER R., FEINER S., JULIER S., MACINTYRE B.: Recent Advances in Augmented Reality, IEEE Computer Graphics and Applications, November/December 2001

[JL04] JACOBS K., LOSCOS C.: Classification of illumination methods for mixed-reality, Proceedings of the Eurographics conference, Grenoble, Sept. 2004.

[FGR93] FOURNIER A., GUNAWAN A. S., ROMANZIN C.: Common Illumination between Real and Computer Generated Scenes. Proc. Graphics Interface. pp. 254-262, May 1993

[WPS*03] WALD I., PURCELL T.J., SCHMITTLER J., BENTHIN C., SLUSALLEK P.: Realtime Ray Tracing and its use for Interactive Global Illumination, Proceedings of Eurographics 2003

[DDM02] DAMEZ C., DMITRIEV K., MYSZKOWSKI K.: Global Illumination for Interactive Applications and High-Quality Animations, Proceedings of Eurographics 2002

[SKS02] SLOAN P.-P., KAUTZ J., SNYDER J.: Precomputed Radiance Transfer for Real-Time Rendering in Dynamic, Low-Frequency Lighting Environments, SIGGRAPH 2002,July, 2002

[Kau04] KAUTZ J.: Hardware Lighting and Shading: A Survey, Computers Graphics Forum 23(1), March 2004, pages 85-112

[PLF*01] PAPAGIANNAKIS G., HOSTE G.L, FONI A., MAGNENAT-THALMANN N.: Real-Time Photo Realistic Simulation of Complex Heritage Edifices, VSMM2001 (Virtual Systems and Multimedia), Conference Proceedings, pp. 218-227, October, 2001

[Can27] CANNON W. B: The james-lange theory of emotion: A critical examination and an alternative theory. American Journal of Psychology, 39:10–124, 1927.

[SS62] SCHACHTER S., SINGER J.: Cognitive, social and physiological determinants of emotional state. Psychol. Rev., 69:379–399, 1962.

[Laz91] LAZARUS R. S.: Emotion and Adaptation. Oxford University Press, New York, 1991.

[Buc80] BUCK R.: Nonverbal behavior and the theory of emotion: The facial feedback hypothesis. Journal of Personality and Social Psychology, 38:811–824, 1980.

[LCK76] LANZETTA J., CARTWRIGHT-SMITH J., KLECK R.: Effects of nonverbal dissimulation on emotion experience and autonomic arousal. Journal of Personality and Social Psychology, 33:354–370, 1976.

[ELF83] EKMAN P., LEVENSON R. W., FRIESEN W. V.: Autonomic nervous system activity distinguishes among emotions. Science, 221:1208–1210, 1983.

[OCC88] ORTONY A., CLORE G. L., COLLINS A.: The Cognitive Structure of Emotions. Cambridge University Press, 1988.

[Eys90] EYSENCK H. J.: Biological dimensions of personality. In L. A. Pervin, editor, Handbook of personality: Theory and research, pages 244–276. New York: Guilford, 1990.

[CM92] COSTA P. T., MCCRAE R. R.: Normal personality assessment in clinical practice: The NEO personality inventory. Psychological Assessment, (4):5–13, 1992.

[EKM04] EGGES A., KSHIRSAGAR S., MAGNENAT-THALMANN N.: Generic personality and emotion simulation for conversational agents. Computer Animation and Virtual Worlds, 15(1):1–13, 2004.

[Piw02] PIWEK P.: An annotated bibliography of affective natural language generation. Technical report, University of Brighton, July 2002.

[MG02] MARSELLA S., GRATCH J.: A step towards irrationality: Using emotion to change belief. In Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems, Bologna, Italy, July 2002.

[JS01] JOHNS M., SILVERMAN B. G.: How emotions and personality effect the utility of alternative decisions: a terrorist target selection case study. In Tenth Conference On Computer Generated Forces and Behavioral Representation, May 2001.

[AKG*99] ANDRE E., KLESEN M., GEBHARD P., ALLEN S., RIST T.: Integrating models of personality and emotions into lifelike characters. In Proceedings International Workshop on Affect in Interactions. Towards a New Generation of Interfaces, 1999.

[BB98] BALL G., BREESE J.: Emotion and personality in a conversational character. In Proceedings of the Workshop on Embodied Conversational Characters, pages 83–84 and 119–121, October 1998.

[KM02] KSHIRSAGAR S., MAGNENAT-THALMANN N.: A multilayer personality model. In Proceedings of 2nd International Symposium on Smart Graphics, pp. 107–115, ACM Press, June 2002.

[Vel97] VELASQUEZ J.: Modeling emotions and other motivations in synthetic agents, In Proceedings of AAAI-97, pages 10–15. MIT Press, 1997.

[Mof95] MOFFAT D.: Personality parameters and programs, In Lecture Notes in Artificial Intelligence: Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents, 1995

II-1: Emotional and Conversational Virtual Humans

Nadia Magnenat-Thalmann

MIRALab, University of Geneva, Geneva, Switzerland

Abstract

In this paper, we present different aspect related to Emotional Conversational Agent. We will describe briefly the Personality and Emotion simulation system based on different research. Next, we will present an short overview of facial animation system, body animation for 3D characters and our face and body animation engine that is used for the study. .

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Three-Dimensional Graphics and Realism]: Virtual Humans, Emotional Model, Facial Animation, Body Animation

1. Overview Embodied Conversational Agents

Over the last years, quite some research has been done to develop systems that can simulate Embodied Conversational Agents. As this paper deals with emotional communicative face and body behaviour for ECAs, we will discuss some recent research related to this topic. One of the most well-known systems that can produce gesture animations from text, is BEAT [CVB01]. BEAT allows animators to input typed text that they wish to be spoken by an animated human figure, and to obtain as output speech and behavioural characteristics. We would also like to mention the MAX system, developed by Kopp and Wachsmuth [KW04]. This system automatically generates gesture animations based on an XML specification of the output. In MAX, the gesture animations are generated procedurally (not from motion captured sequences). The work of Hartmann et al. [HMP02] provides for a system to automatically generate gestures from conversation transcripts using predefined key-frames.

2. Personality and Emotion Simulation

An important control mechanism for ECAs is a personality/emotion simulator. Personality and emotions have a significant effect on how one perceives, thinks and acts (see Figure 1). In this section, we will give a short overview of the different existing techniques for including emotions into perception and reasoning. After this section, we will give some examples of how personality and emotion can play a role in expression.

When discussing theory and simulation of personality and emotion, we have to address work done in both Computer Science and Psychology. Psychology tries to discover the nature of emotions and personality. It describes the structure of each of these notions, how it can be attributed and what their effect is on human behaviour. Computer Science tries to simulate the effects that personality and emotion has on human beings and use it to make Human Computer Interaction (HCI) more natural. The research that is a part of HCI and that tries to use emotion -and in a lesser way personality- to increase naturalness of HCI, is called Affective Computing [Pic97]. We will first discuss some general features of personality and emotion. Then we will discuss how personality and emotion from a part of our behaviour.

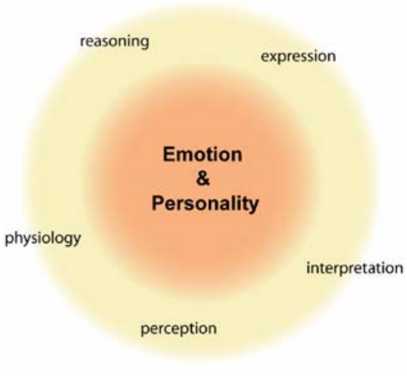
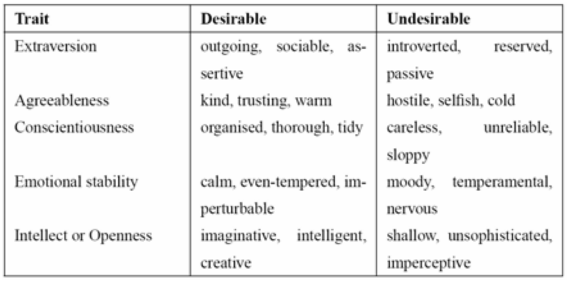


Figure 1: *Personality and emotion and their link with ECA system parts.*

2.1. Personality

The study of personality investigates how individuals differ from each other. The most commonly accepted form of personality study is based on trait theory. Personality traits are relatively stable dispositions that give rise to characteristic patterns of behaviour [Ham99]. Although trait-based personality models have been criticized in the past [Mis68], they are still one of the most popular representations of personality. Among personality researchers, there is still debate on how many traits are needed to provide a comprehensive description of personality. For example, Eysenck proposes three traits of personality: extraversion, neuroticism and psychoticism [Eys90, Eys91], whereas Cattell et al. advocates 16 traits [CET70]. However, the most widely accepted theory is a structure consisting of five factors (also called the Big Five). One of the advantages of the Big Five framework is that it can assimilate other structures. An empirical demonstration of this property is given by Goldberg and Rosolack [GR94] by integrating Eysenck's threefactor system into the Big Five. Each of the five dimensions is defined by its desirable and undesirable qualities. Figure 2 summarizes the Big Five model and the desirable and undesirable qualities of each trait.



Trait	Desirable	Undesirable
Extraversion	outgoing, sociable, assertive	introverted, reserved, passive
Agreeableness	kind, trusting, warm	hostile, selfish, cold
Conscientiousness	organised, thorough, tidy	careless, unreliable, sloppy
Emotional stability	calm, even-tempered, imperturbable	moody, temperamental, nervous
Intellect or Openness	imaginative, intelligent, creative	shallow, unsophisticated, imperceptive

Figure 2: *The Big Five personality traits and their desirable and undersirable qualities [Ham99]*

Not all personality researchers believe that personality is a static set of traits that can be described independently of behaviour and situations. In fact, Mischel and Shoda [MS95, MS98] have developed a personality theory, that accounts for variability patterns across different situations. Although the trait-based approach to modelling personality does have some disadvantages, it is still quite useful for ECA development, because the concept of personality traits can be easily translated into a computational model (for an example, see Johns and Silverman [JS01]). Furthermore, a lot of resources (such as five factor personality tests) are available, so that any research performed can be easily evaluated and adapted.

2.2. Emotion

The concept of emotion has been widely researched in the psychology field. Many approaches and theories exist, but according to Cornelius [Cor96], they can be broadly organised in four different theoretical perspectives: the Darwinian, Jamesian, cognitive, and social constructivist perspectives.

The central organizing idea of the Darwinian perspective is the notion that emotions are phenomena that have evolved from important survival functions. This means that we should see more or less the same emotions (and expressions) in all humans. Theory and research from the Jamesian perspective views emotions from the perspective that it would be impossible to have emotions without bodily changes and bodily changes always come first. The cognitive approach to the study of emotions started with the work of Arnold [Arn60]. In this approach, the central assumption is that thought and emotions are inseparable. More specifically, all emotions are seen within this perspective as being dependent on appraisal, the process by which events in the environment are judged as good or bad for us. Finally, Cornelius identifies the social constructivist approach to emotion study. One of the first emotion studies based on this approach was performed by Averill [AVEve80] and Harré [Har86]. Social constructivists believe that emotions are cultural products that owe their meaning and coherence to learned social rules. Culture, for social constructivists, plays a central role in the organization of emotions at a variety of levels.

In each of these perspectives, it is generally assumed that an emotional state can be viewed as a set of dimensions. The number of these dimensions varies among different researchers. Ekman [Ekm82] has identified six common expressions of emotion: fear, disgust, anger, sadness, surprise and joy, Plutchik [Plu80] proposes eight, and in the OCC appraisal model [OCC88], there are 22 emotions. Our system is based on an emotion representation called the activation-evaluation space [Sch54], which defines emotions along two axes on a circle (see Figure 3), where the distance from the centre defines the power of the emotion. This emotion space allows us to easily map the different types of idle motions. Additionally, different discrete emotions can be placed on the disc [CDCS*00], which provides for a possibility to link the activation-evaluation model with different multidimensional emotion models, such as the OCC emotions or Ekman's expressions.

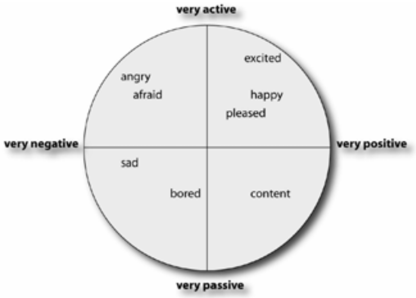


Figure 3: *Activation-evaluation emotion disc*

3. Dialogue Systems

One of the first attempts to create a system that interacts with a human through natural language dialogue was ELIZA [CBB*99]. It used pattern-matching techniques to extract information from a sentence. ELIZA had no sophisticated dialogue model, it just had a simple trigger-reaction mechanism. Although the approach is very limited and dialogue systems have evolved, it is still popular, see for example AIML [AIML]. In this section, we will give a brief survey of the state of the art in dialogue systems. We will discuss various types of dialogue management and finally we discuss a general theory of dialogue: the information state theory.

This subsection presents the different methods that exist for modeling dialogue. A more complete survey is given in [GCAS97]. In general, there are three main approaches to dialogue modeling: dialogue grammars, plan-based dialogue management and collaborative dialogue.

For describing the structure of a dialogue, one of the first methods was to use dialogue grammars. Dialogue grammars define sequences of sentences in a dialogue. Often, such a grammar describes the whole dialogue from beginning to end, but recent approaches used these grammars for describing often occurring sequences in a dialogue, such as question-answer pairs. Constructing such a grammar can be done by using Chomsky-like rules or a finite-state machine. The latter is especially suited when the dialogue structure is the same as the task structure, which is: in every state the user has certain choices and these choices correspond to state transitions. Because the system always takes the initiative in this approach, it can very well anticipate on the user’s response, since there are only a few choices presented to the user.

Plan-based approaches can model more complex dialogues than dialogue grammars. The idea of a plan-based approach for the listener is to discover the underlying plan of the speaker and respond properly to this. If such a plan is correctly identified, this approach can very well handle indirect speech acts. However, a plan-based approach does not explain why dialogue takes place; the dialogue manager has no explicit goal.

A collaborative approach views a dialogue as collaboration between partners in order to achieve a mutual understanding of the dialogue. This commitment of both participants is the reason that dialogues contain clarifications, confirmations and so on. Every partner has certain beliefs about the other partner and uses these beliefs to respond to him/her. Also, two partners can share mutual beliefs: statements that both participants regard to be true. This concept of mutual belief is present in the model of conversational agency in [Tra96] as an extension to the BDI model [Bra87]. BDI stands for belief, desire, intention, which are the three possible kinds of statements that can occur in an agent’s state. A lot of BDI systems use some kind of extension of logic to express beliefs, desires and intentions. This is called a BDI logic (for an example of this, see [Woo00]). Applications that use the BDI model to develop a conversational agent are described in [SBP97].

Altogether, several approaches to developing dialogue systems can be compared through the information state theory, which was defined in the TRINDI project [LT00]. In fact, the information state theory is an attempt to develop a model for best practice in the development of the dialogue management component of a spoken dialogue system. Key to this approach is to identify the relevant aspects of information in dialogue, how they are updated and how updating processes are controlled. This framework can then allow comparison to determine empirically which the best practice is.

In the next section Facial Animation System, we discuss different approaches to simulate facial movements on a 3D face model. Also, we should present some methods to parameterize these approaches so that the control mechanism of the face (for example FACS or MPEG-4) is independent of the 3D model. However, a higher level control is required when one wants to control a face as a part of an ECA. One of the main output formats of an ECA is speech.

Therefore a mechanism is required that can convert this speech into an accurate facial animation signal. Additionally, emotional expressions need to be blended with a speech signal, as well as other face motions, such as head movements, gaze and eye blinking. More specifically, Ekman [Ekm82] defines several different groups of facial expressions during speech: emblems, emotion emblems, conversational signals, punctuators, regularos, manipulators and affect displays. These groups serve either as communicative signals or as movements corresponding to a biological need (eye blinking, wetting the lips, and so on).

4. Facial Animation System

4.1. Parameterization

The first step in animating faces, like most other 3D objects, is to devise a way of describing the facial deformation and to use this deformation description to define key frames of animation. Facial Deformation is the process of modifying or changing the shape of a given facial mesh by displacing a number of vertices from their neutral position, usually to affect small local regions, in order to portray a change in facial expression. The description of deformation can be done most optimally by devising a parameter set or in other words “parameterizing” the facial deformation. Of course, in the very beginning, parameterization was not used. The initial method of animating faces was to setup key-frames by manipulating the facial mesh at a low level, by actually displacing vertices and triangles for each key frame of animation [Par72]. Soon parameterized models were developed that allowed the designers to quickly change facial expressions by manipulating a set of only a few parameters controlling different parts of the face. The facial deformation techniques depend on the underlying parameterization scheme to a great extent. Not all the deformation techniques can be used with all parameterization schemes and vice versa.

Amongst the various facial control parameterization schemes, the earliest and probably the most commonly used is the Facial Action Coding System (FACS, [EF78]). FACS was originally indented only for describing all possible visually distinguishable facial movements or *actions*. The system describes the most basic facial muscle actions and their effect on facial expressions. Though the FACS was not formulated with computer facial animation in mind, many animation schemes are based on FACS or are inspired from FACS as the facial control parameters. FACS is used as a way to control facial movements by specifying the muscle actions needed to achieve desired expression changes. Since FACS is based on muscle actions, the corresponding facial deformation techniques are based on simulating the characteristics of the facial muscles. The facial animation system developed by Kalra *et al* [KMMT91] used the Minimal Perceptible Action (MPA) as basic facial motion parameter. This scheme was inspired from FACS, but provided more detailed and asymmetric facial movements, and supported global head movements such as “nod head”. Recently, MPEG-4 facial animation standard introduced Facial Animation Parameters (FAP) as a standard parameter set for animating synthetic 3D faces [MPEG4, Ost98]. Since then it has been widely adopted for various facial animation systems.

4.2. Facial deformation

For facial deformation, one of the earliest approaches employed was shape interpolation [Par72, BL85]. In this approach, facial animation is obtained by linear or non-linear interpolation functions applied to vertex positions of various facial poses. No particular parameter set was used. In parametric models, the facial mesh is manipulated through a set of parameters [Par82]. The interpolation functions are applied to the parameters rather than individual vertex position, and facial deformation is obtained for each frame from these interpolated parameters. However, the design of the parameters set is based on hard-wiring the vertices for manipulating a part of the face, which makes the model dependent on the facial topology. The advantages of parametric deformation over the shape interpolation deformation area many ease of design, storage and manipulation of animation.

Amongst the physics based models using muscle based parameter set (mainly FACS); the earliest attempt was by Platt and Badler [PB81]. They used forces applied to elastic meshes through muscle arcs for generating realistic facial expressions. Waters [Wat87] developed a “vector muscle” model. A muscle was defined using a vector field direction, an origin, and an insertion point. The field extent defined the influence of the muscle activity on the facial region. The mouth sphincter muscle was modeled as a simplified parametric ellipsoid. The model proposed by Terzopoulos and Waters [TW90] had detailed anatomical structure and dynamics of the human face, consisting of three-layers corresponding to skin, fatty tissue, and muscle tied to bone. Elastic spring elements were used to simulate facial deformation at each layer. During last years, a lot of researches as been done in order to extend and optimize these approaches but the concept remained the same ones.

In this next section we present the concept developed in the work done by Garchery S. [Gar04] and Kshirsagar S. [Ksh03] at MIRALab in facial deformation and animation system based on MPEG-4 parameters.

4.3. Case study: MPEG-4 facial animation engine

If you are not familiarly with MPEG-4 parameterization for facial animation, please refer to chapter 16 in Magnenat-Thalmann & Thalmann [MT04] who describe all principal aspect of MPEG-4 for face application.

We would like to present briefly a generic approach developed at MIRALab in order to compute automatically the influence of each Facial Animation Parameters on the face model. The main idea is to keep coherence and a maximum of simplicity in order to adapt this approach to different platforms or environments, but in the same time also to be able to produce realistic expressions. Exactly the same process is used for each FAP (Facial Animation Parameter) in order to define the influence area. In order to obtain a maximum of simplicity, we have developed an approach to compute deformations from the minimum information: simply FDP (Facial Definition Parameters) information.

4.3.1. Computation description

The main problem is to find a correct definition of the influence area for each FAP according to its neighboring vertices (i.e. define a correct influence area according to deformation needed). We propose an approach based on the following process:

Compute the distribution of FDPs on the model and a relation of distance between them. Then, for each vertex of the mesh:

- Define which control point is able to influence it
- Define the ratio of influence of each control point

A 3D face model is composed of different meshes for eyes, teeth, tongue and skin. The skin mesh is mainly defined by the holes for the eyes and the mouth. Also, often a face model has a vertex distribution that is not uniform over the mesh. In order to develop a model-independent approach we have to take into account these specificities, and then we should define an appropriate distance measurement. A measurement based on Euclidian distance is efficient to manage the variations in mesh density but it does not take into account problems like holes in the model.

A measurement based on the topology like the number of edges between vertices takes the holes into account, but it is not efficient for the mesh density variations.

We propose to use a metric based on both aspects: Euclidian distance and mesh topology. The metric is computed following this rule: “the distance is equal to the sum of the edge distances along the shortest path between two vertices”. Using this metric, we are able to manage in the same time, holes and mesh density variation on the face model.

Our approach is based on the definition of a list of influencing feature points for each vertex in the face mesh. Initially, all of the feature points are in the influence list. First, we find the closest feature point to the vertex (according to the previously defined metric). Then, we remove all the feature points from the list that are in the plane perpendicular to the vector between the vertex and the selected feature point (see Figure 4). Then, we select the next closest feature point in the remaining list and we apply the same procedure until all feature points have been taken into account.

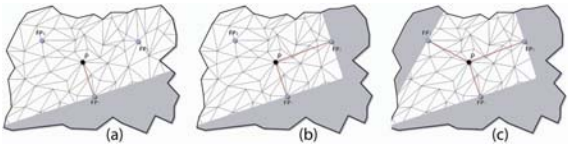


Figure 4: *Step by step definition of feature point sample.*

When the list of influencing feature points is established, we compute the influence of each of them to the current vertex by using a rule based on the metric explained above. If *P* is a vertex of the mesh, we compute a balanced sum *d*:

$$d = \frac{\sum_{i=1}^{n} d_i * \cos\theta_i}{\sum_{i=1}^{n} \cos\theta_i}$$

where *n* equals the number of influencing feature points for *P*, θ_i the angle between *P* and the feature point and d_i the distance between the feature point and the vertex. The weight associated to a specific feature point for *P* is computed as:

$$W_{i,P} = \sin\left(\frac{\pi}{2}\left(1 - \frac{d_i}{d}\right)\right)$$

By applying this weight computation for each vertex of the mesh, we obtain for each feature point a list of the vertices that are influenced by it, with an associated weight with the respect of mesh continuity.

This approach present different advantages like taking account of overlapping regions and the diversity of the mesh (variation of density and holes) and is independent of number and repartition of feature points.

4.3.2. Application to the face by defining simple constraint

This generic approach is applied on the 3D face model 3 times: one time for each direction of deformation. When we look at the FAP repartition from directional point of view, we can see a big variation in the feature point’s density on the face (see Figure 5). In order to take into account this diversity and produce a realistic deformation, we compute a different influence area according to each displacement direction. We obtain then 3 different set of vertices influenced for each feature point for each region. We use this information during the real time animation in order to deform the mesh according to feature point’s displacements. This information can easily be represented in the FaceDefMesh format and be used in an MPEG-4 compliant face system.

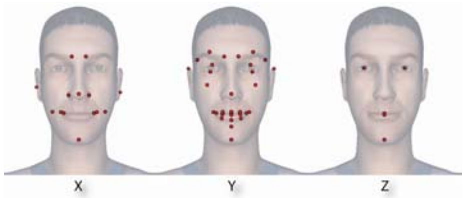
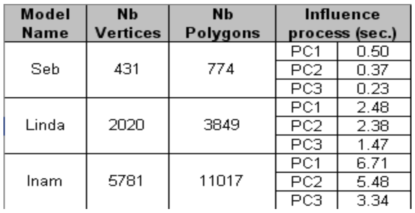


Figure 5: *Directional repartition of Facial Animation Parameters*

This initialization step of influence computation is done only once or is already saved in the MPEG-4 format. This approach is computationally very efficient (see Figure 6). We only need a few seconds to compute all this information for a model composed with more than 10K polygons on a standard PC.



Model Name	Nb Vertices	Nb Polygons	Influence process (sec.)	
Seb	431	774	PC1	0.50
			PC2	0.37
			PC3	0.23
Linda	2020	3849	PC1	2.48
			PC2	2.38
			PC3	1.47
Inam	5781	11017	PC1	6.71
			PC2	5.48
			PC3	3.34

Figure 6: *Computation time on different face model. PC1 = AMD Athlon 1.7+, PC2 = PIV 2 Ghz, PC3 = PIV 3 Ghz.*

The advantages of our approach are multiple: we find a quick way to simplify the number of influent point; the same computation applies to all vertices independent of the part of face. Therefore we can add or remove feature points easily without changing the computation process. This approach works not only with MPEG-4 but with any feature point-based deformation approach (like information from optical capture system). The next picture present some expressions computed in real time with a 3D model composed with 14.500 polygons.

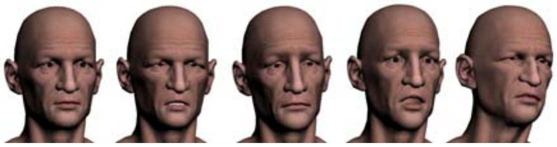
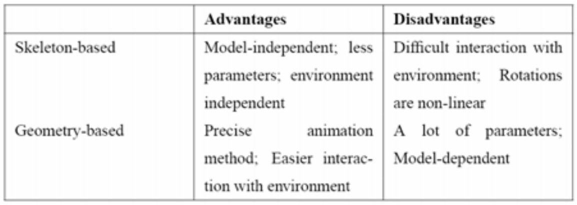


Figure 7: *Facial expression computed with real-time MIRA-Lab approach*

5. Body Animation

5.1. Skeleton-Based Animation

For producing real-time body animations, it is common practice to work on the underlying skeleton in stead of the model itself (please see Figure 8). Such an animation can then be defined by a set of joint rotations and translations. There are currently two well known standards for body animation: MPEG-4 and VRML H-Anim. We will now shortly discuss each of these standards.



	Advantages	Disadvantages
Skeleton-based	Model-independent; less parameters; environment independent	Difficult interaction with environment; Rotations are non-linear
Geometry-based	Precise animation method; Easier interaction with environment	A lot of parameters; Model-dependent

Figure 8: *Advantages and disadvantages for two different animation approaches*

The H-Anim specification of animatable humanoids forms a part of the VRML language. H-Anim defines a hierarchy of joints, with as a root node the so called HumanoidRoot. The latest upgrade of the standard (H-Anim 200x), defines four levels of articulation (0-3). In Figure 9, the joint rotation centers for Level of Articulation One are depicted. Since the H-Anim 200x standard is quite new, a lot of recent work is based on the H-Anim 1.1 standard.

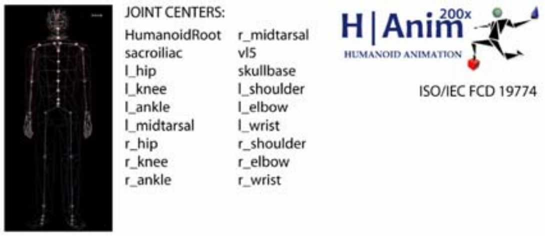


Figure 9: *The H-Anim 200x standard, level of articulation One*

The part of MPEG-4 that addresses the representation and coding of synthetically and naturally generated audio-visual information is called SNHC (Synthetic/Natural Hybrid Coding). The Animation Framework extension (AFX - pronounced 'effects') provides enhanced visual experiences in synthetic MPEG-4 environments. The framework defines a collection of interoperable tools that allow for example to add higher level descriptions of animations (such as inverse kinematics), scalability and compression techniques.

Similar to the facial animation definition in MPEG-4, two sets of parameters are defined for describing and animating the body: the Body Definition Parameter (BDP) set, and the Body Animation Parameter (BAP) set (see also Figure 10). The BDP set defines the set of parameters to transform the default body to a customized body with its body surface, body dimensions, and (optionally) texture. The Body Animation Parameters (BAPs), if correctly interpreted, will produce reasonably similar high level results in terms of body posture and animation on different body models, without the need to initialize or calibrate the model.

Upon construction, the Body object contains a generic virtual human body with the default posture. This body can already be rendered. It is also immediately capable of receiving the BAPs from the bitstream, which will produce animation of the body. If BDPs are received, they are used to transform the generic body into a particular body determined by the parameters contents. No assumption is made and no limitation is imposed on the range of motion of joints.

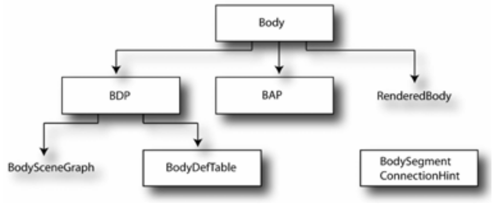


Figure 10: *Definition and animation of virtual body in the MPEG-4 Standard*

5.2. Representation of Skeleton Posture

Given that a 3D Humanoid model is being controlled through a skeleton motion, it is important to choose the right representation for the transformations of the skeleton joints. The most basic representation is to define a transformation matrix that contains a translation and a rotation component, where the rotation is defined as a 3×3 matrix. Transforming a skeleton joint becomes tricky when looking at the rotation component. A rotation involves three degrees of freedom (DOF), around the x, y and z axis, whereas a rotation matrix defines 9 (as a 3 × 3 matrix). A matrix can only represent a rotation if it is orthonormalised. If this condition is violated, arbitrary rotations can occur. This means that during animation, additional operations are required to ensure the orthonormality of the matrix, which is computationally intensive. Furthermore, rotation matrices are not very well suited for rotation interpolation. Therefore, other representations of joint rotations have been proposed. For a more detailed discussion of each of the representations, we refer to Bobick [Bob98].

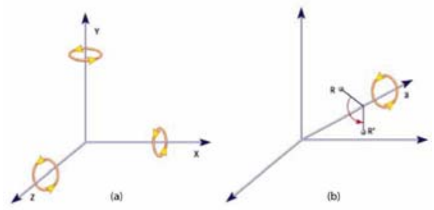


Figure 11: *Two different angular representations of rotations: (a) Euler angles (b) Axis angles*

When a rotation is represented as three rotations around the three coordinate axes, it is called the Euler Angle representation (see Figure 11). An Euler angle can be written as (rx, ry, rz), meaning: Rotate a point counterclockwise around the x axis by rx degrees, followed by a rotation around the y axis by ry degrees, followed by a rotation around the z axis by rz degrees. The Euler angle representation is quite efficient since it uses three variables (angles) to define three degrees of freedom. However, interpolation can be computationally expensive, since it requires numerical integration. Furthermore, Euler Angles have a Gimbal lock problem (or: the loss of one degree of freedom), that occurs when a series of rotations at 90 degrees are performed. Due to the alignment of the axes, these rotations can cancel out each other.

Another representation of rotation is called Axis Angle (see also Figure 11). This representation defines an arbitrary axis and a rotation around this axis. Although this representation is quite efficient, it still requires a lot of computational efforts for interpolation.

Quaternions are also a common way to represent rotations. Basically, a quaternion is a four-dimensional complex number that can be used to represent orientations in three dimensions. A quaternion is popularly written as:

$$w + xi + yj + zk$$

Quaternions can be interpolated using a method called SLERP (Spherical Linear intERPolation) [Bob98]. This is easier to use than rotation matrix interpolation and suitable for real-time application. A disadvantage of quaternions is that they are difficult to visualize and that they are completely unintuitive.

Finally, rotations can be represented using the exponential map. Alexa [Ale02] has described a method using the exponential map to allow for transformations that are performed completely in the linear domain, thus solving a lot of problems that the previous methods are suffering from. Rotations are represented by a skew-symmetric matrix. For every real skew-symmetric matrix, its exponential map is always a rotation matrix (see for example Chevalley [Che46]). Conversely, given a rotation matrix R, there exists some skew-symmetric matrix B such that R = eB. The skew-symmetric matrix representation of a rotation is very useful for motion interpolation [PR97], because it allows performing linear operations on rotations. A three-dimensional real skew-symmetric matrix has the following form:

$$B = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}$$

Such an element can also be represented as a vector r where:

$$r = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The exponential map representation r represents a rotation of $\theta = \sqrt{a^2 + b^2 + c^2}$ degrees around axis r. The exponential of a matrix B is defined by Rodrigues' formula:

$$e^B = I_3 + \frac{\sin\theta}{\theta}B + \frac{(1-\cos\theta)}{\theta^2}B^2$$

Similarly, methods exist that define how to calculate a determination of the (multivalued) matrix logarithm. For example, Gallier and Xu [GX02] present methods to calculate exponentials and logarithms, also for matrices with higher dimensions.

Although the exponential map representation of orientation does allow for easy manipulation and interpolation, there are singularities in the domain. A log map can map each orientation to an infinite number of points, corresponding to rotations of $2n\pi + \theta$ about axis v and $2n\pi - \theta$ about axis -v for any $n \in \mathbb{N}$ [Gra98]. Consequently, measures have to be taken to ensure that these singularities do not interfere with the interpolation (please see the paper by Grassia [Gra98] for a more detailed discussion).

5.3. Deforming the Body from Skeleton Posture

Skeleton-driven deformation (SDD), a classical method for the basic skin deformation is perhaps the most widely used technique in 3D character animation. In research literature, an early version was presented by Magnenat-Thalmann et al. [MLT88], who introduced the concept of Joint-dependent Local Deformation (JLD) operators to smoothly deform the skin surface. This technique has been given various names such as Sub-Space Deformation (SSD), linear blend skinning, or smooth skinning. Several attempts have been made to overcome the limitation of geometric skin deformation by using examples of varying postures and blending them during animation. Pose space deformation [LCF00] approaches the problem by using artistically sculpted skin surfaces of varying posture and blending them during animation. More recently, Kry et al. [KJE02] proposed an extension of that technique by using principal component analysis (PCA), allowing for optimal reduction of the data and thus faster deformation. Sloan et al. [SRC01] have shown similar results using RBF for blending the arm models. Their contribution lies in that they make use of equivalent of cardinal basis function. Allen et al. [ACP02] present another example-based method for creating realistic skeleton-driven deformation. More recently, Mohr et al. [MG03] have shown the extension of the SDD by introducing pseudo joints. Finally, Magnenat-Thalmann et al. [MCSP04] propose an extension of the SDD that overcomes the undesirable effect of vertex collapsing (see Figure 12).

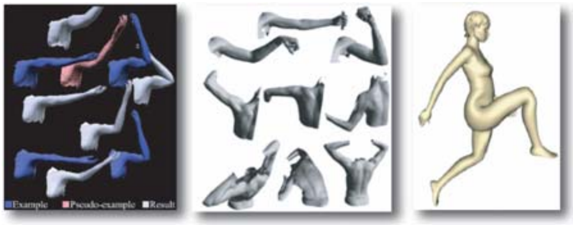


Figure 12: *Results of skinning algorithms by Sloan et al, Allen et al and Magnenat-Thalman et al.*

5.4. Body inside ECA System

In order to link the body animations with ECA systems, a translation is required from the higher level definitions of ECA gesture synthesizers and the low-level animations. Generally, such motions are produced procedurally, although this results sometimes in mechanical motions. Recent work from Stone et al. [SDO04] describes a system that uses motion capture data to produce new gesture animations. The system is based on communicative units that combine both speech and gestures. The existing combinations in the motion capture database are used to construct new animations from new utterances. This method does result in natural-looking animations, but in order to provide for a wide range of motions, a coherent performance from the motion captured person is required. Also, the style and shape of the motions are not directly controllable, contrary to procedural animation methods. Especially when one desires to generate motions that reflect a certain style, emotion or individuality, a highly flexible animation engine is required that allows for a precise definition of how the movement should take place, while still retaining the motion realism that can be obtained using motion capture techniques. The EMOTE model [CCZB00] aims to control gesture motions using effort and shape parameters. As a result, gestures can be adapted to express emotional content or to stress a part of what is communicated.

In order to integrate the idle motions and the gesture animations, we use the blending library that was developed in our earlier work [EMT04]. This library allows us to perform weighted animation blending operations in the exponential map space. Additionally, a set of modifiers is provided that allows scaling, flipping and stretching of animations, among others. Figure 13 shows the general process of the animation synthesis and blending. The idle motion engine is running continuously, therefore providing the ECA with continuous idle motions. Gesture animations (that are for example generated by a text-to-gesture system such as BEAT) are blended in on-the-fly.

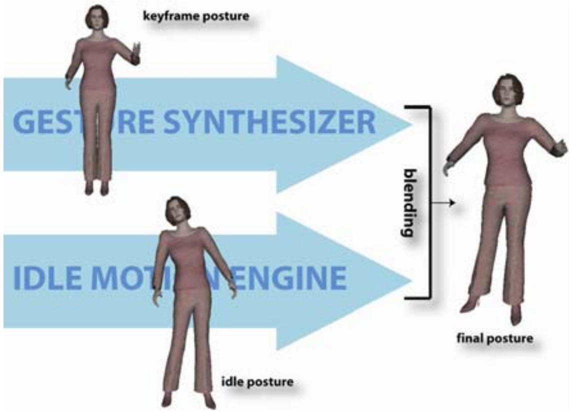


Figure 13: *Blending of idle posture and gesture, resulting in the final posture*

References

[AIML] AIML webpage. http://www.alicebot.org.

[Ale02] ALEXA M.: Linear combination of transformations. In SIGGRAPH 2002, pages 380–387, 2002.

[ACP02] ALLEN B., CURLESS B. POPOVIC Z: Articulated body deformation from range scan data. In Proceedings SIGGRAPH 2002, pages 612–619. Addison-Wesley, 2002.

[Arn60] ARNOLD M. B.: Emotion and personality. Columbia University Press, New York, 1960.

[Ave80] AVERILL J. R.: A constructivist view of emotion. In Emotion: Theory, research and experience, Plutchik R., Kellerman H., (Eds.), vol. 1. Academic Press, New York, 1980, pp. 305–339.

[Bra87] BARTMAN M.E.: Intentions, Plans, and Practical Reason. Harvard University Press, 1987.

[BL85] BERGERON P, LACHAPELLE P (1985) Controlling facial expression and body movements in the computer generated short ‘tony de peltrie’. In SIGGRAPH 85 Tutorial Notes (1985), ACM Press.

[Bob98] BOBICK N.: Rotating objects using quaternions. Game Developer, February 1998.

[CVB01] CASSELL J., VILHJ’ALMSSON H., BICKMORE T.: BEAT: the Behavior Expression Animation Toolkit. In Proceedings of SIGGRAPH, pages 477–486, 2001.

[CBB*99] CASSELL J:, BICKMORE T, BILLINGHURST M., CAMPBELL L, CHANG K., VILHJ'ALMSSON H, YAN H.: Embodiment in conversational interfaces: Rea. In Proceedings of the CHI'99 Conference, pages 520–527, 1999.

[CCZB00] CHI D., COSTA M., ZHAO L., BADLER N.: The emote model for effort and shape. In SIGGRAPH 2000 (July 2000), pp. 173-182.

[CET70] CATELL B, EBER H., TATSUOKA M.: Handbook for the Sixteen Personality Factor Questionnaire (16PF). Institute for Personality and Ability Testing, Champaign, IL, 1970.

[Che46] CHEVALLEY C.. Theory of Lie Groups. Princeton University Press, New York, 1946.

[Cor96] CORNELIUS R. R.: The science of emotion. Research and tradition in the psychology of emotion. Prentice-Hall, Upper Saddle River (NJ), 1996.

[CDCS*00] COWIE R., DOUGLAS-COWIE E., SAVVIDOU S., MCMAHON E., SAWEY M., SCHRÖDER M.: Feeltrace: An instrument for recording perceived emotion in real time. In ISCA Workshop on Speech and Emotion (Northern Ireland, 2000), pp. 19–24.

[EMT04] EGGES A., MOLET T., MAGNENAT THALMANN N.: Personalised real-time idle motion synthesis. In Proceedings of the 12th Pacific Graphics Conference, pages 121- 130, October 2004.

[Ekm82] EKMAN P.: Emotion in the human face. Cambridge University Press, New York, 1982.

[EF78] EKMAN P, FRISEN WV (1978) Facial Action Coding System, Investigator's Guide Part II, Consulting Psychologists Press Inc.

[Eys90] EYSENCK J.: Biological dimensions of personality. In L. A. Pervin, editor, Handbook of personality: Theory and research, pages 244–276. New York: Guilford, 1990.

[Eys91] EYSENCK J. : Dimensions of personality: 16, 5 or 3 criteria for a taxonomic paradigm. Personality and Individual Differences, 12:773–790, 1991.

[GX02] GALLIER J., XU D.: Computing exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. International Journal of Robotics and automation, 17(4), 2002.

[Gar04] GARCHERY S.: Animation Faciale Temps-Reel Multi Plates-Formes. PhD thesis, MIRALab, University of Geneva, 2004.

[GCAS97] GAVIN E., CHIRCHER E., ATWELL, SOUTER C.: Dialogue management systems: a survey and overview. Technical report, University of Leeds, February 1997.

[GR94] GOLDBERG L.R., ROSOLACK T.K.: The big five factor structure as an integrative framework: An empirical comparison with eysenck's p-e-n model. In C. F. Halverson Jr., G. A. Kohnstamm, and R. P. Martin, editors, The Developing Structure of Temperament and Personality from Infancy to Adulthood. Lawrence Erlbaum, New York, 1994.

[Gra98] GRASSIA F.S.: Practical parameterization of rotations using the exponential map. Journal of Graphics Tools, 3(3):29–48, 1998.

[Ham99] HAMPSON S. State of the art: Personality. The Psychologist, 12(6):284–290, June 1999.

[Har86] HARRÉ R. (Ed.): The social construction of emotions. Basil Blackwell, Oxford, 1986.

[HMP02] HARTMANN B., MACNINI M., PELACHAUD C.: Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis. In Computer Animation 2002, pages 111–119, 2002.

[JS01] JOHNS M. SILVERMAN B.G.: How emotions and personality effect the utility of alternative decisions: a terrorist target selection case study. In Tenth Conference On Computer Generated Forces and Behavioral Representation, May 2001.

[KMMT91] KARLA P, MANGILI A., MAGNENAT THALMANN N, THALMANN D. : (1991) SMILE: A Mult-layered Facial Animation System, Proc. IFIP WG 5.10, Tokyo, Japan (ed TL Kunii), pp 189-198

[KW04] KOPP S. WACHSMUTH I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds, 15(1):39–52, 2004.

[KJE02] KRY P:G:, PAI. D.: Eigenskin: Real time large deformation character skinning in graphics hardware. In ACM SIGGRAPH Symposium on Computer Animation, pages 153–159, 2002.

[Ksh03] KSHIRSAGAR S.: Facial Communication. PhD thesis, MIRALab, University of Geneva, 2003.

[LT00] LARSSON S., TRAUM D.: Information state and dialogue management in the TRINDI dialogue move engine toolkit. Gothenburg papers in computational linguistics, April 2000.

[LCF00] LEWIS J.P., CORDNER M., FONG N.: Pose space deformations: A unified approach to shape interpolation and skeleton-driven deformation. In Proceedings SIGGRAPH 2000, pages 165–172, 2000.

[MCSP04] MAGNENAT-THALMANN N, CORDIER F. SEO H. PAPAGIANNAKIS G.: Modeling of bodies and clothes for virtual environments. In Cyberworlds, pages 201–208, July 2004.

[MLT88] MAGNENAT-THALMANN N., LAPERRIRE R, THALMANN D.: Joint-dependent local deformations for hand animation and object grasping. In Proceedings Graphics Interface, pages 26–33, 1988.

[MT04] MAGNENAT-THALMANN N., THALMANN D. (2004): Handbook of Virtual Human, *Eds Wiley & Sons, Ltd.*, publisher, ISBN: 0-470-02316-3.

[Mis68] MISCHEL W.: Personality and Assessment. Wiley, New York, 1968.

[MY95] MISCHEL W., SHODA Y.: A cognitive-affective system theory of personality: reconceptualising situations, dispositions, dynamics and invariance in personality structure. Psychological Review, 102:246–268, 1995.

[MY98] MISCHEL W., SHODA Y.: Reconciling processing dynamics and personality dispositions. Annual Review of Psychology, 49:229–258, 1998.

[MG03] MOHR A., GLEIDHER M.: Building efficient, accurate character skins from examples. In Proceedings SIGGRAPH 2003, pages 165–172, 2003.

[MPEG4] MPEG-4 Specification of MPEG-4 standard, Moving Picture Experts Group. http://www.cselt.it/mpeg.

[OCC88] ORTONY A., CLORE G. L., COLLINS A.: The Cognitive Structure of Emotions. Cambridge University Press, 1988.

[Ost98] OSTERMANN J. (1998) Animation of synthetic faces in mpeg-4. Computer Animation '98 Held in Philadelphia, Pennsylvania, USA.

[Par72] PARKE F.I. (1972) Computer Generated Animation of Faces. Proc. ACM annual conference

[Par82] PARKE F.I. (1982) Parameterized models for facial animation. IEEE Computer Graphics and Applications 2, 9 (November 1982), 61–68.

[PR97] PARK F.C., RAVANI B.: Smooth invariant interpolation of rotations. ACM Transactions on Graphics, 16(3):277–295, July 1997.

[Pic97] PICARD R.W. : *Affective computing*. MIT Press, Cambridge, MA, 1997.

[PB81] PLATT S., BADLER N. (1981) Animating facial expression. Computer Graphics 15, 3 (1981), 245–252.

[Plu80] PLUTCHIK R.: Emotion: A psychoevolutionary synthesis. Harper & Row, New York, 1980.

[SBP97] SADEK M.D., BRETIER P., PANAGET F.: ARTIMIS: Natural dialogue meets rational agency. In M. E. Pollack, editor, Proceedings 15th International Joint Conference on Artificial Intelligence, pages 1030–1035. Morgan Kaufmann Publishers, 1997.

[Sch54] SCHLOSBERG H.: A scale for judgement of facial expressions. Journal of Experimental Psychology 29 (1954), 497–510.

[SDO04] STONE M., DECARLO D., OH I., RODRIGUEZ C., STERE A., LEES A., BREGLER C.: Speaking with hands: Creating animated conversational characters from recordings of human performance. In SIGGRAPH 2004 (2004), pp. 506-513.

[SRC01] SLOAN P.P., ROSE C., COHEN M.: Shape by example. In Symposium on Interactive 3D Graphics, March 2001.

[TW90] TERZOPOULOS D., WATERS K. (1990) Physically-based facial modelling, analysis, and animation. Journal of Visualization and Computer Animation 1, 2 (August 1990), 73–80. West Sussex, England.

[Tra96] TRAUM D.R. : Conversational agency: The TRAINS-93 dialogue manager. In Proceedings of the Twente Workshop on Langauge Technology: Dialogue Management in Natural Language Systems (TWLT 11), pages 1–11, 1996.

[Wat87] WATERS K. (1987) A muscle model for animating three-dimensional facial expression. Computer Graphics (Proceedings of SIGGRAPH 87) 21, 4 (July 1987), 17–24. Anaheim, California.

[Woo00] WOOLDRIDGE M.: Reasoning about rational agents. MIT Press, 2000.

II-2: Vision Based 3D Tracking and Pose Estimation for Mixed Reality

P. Fua and V. Lepetit

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Computer Vision Laboratory
CH-1015 Lausanne, Switzerland

Abstract

Mixed Reality applications require accurate knowledge of the relative positions of the camera and the scene. When either of them moves, this means keeping track in real-time of all six degrees of freedom that define the camera position and orientation relative to the scene, or, equivalently, the 3D displacement of an object relative to the camera.
Many technologies have been tried to achieve this goal. However, Computer Vision is the only one that has the potential to yield non-invasive, accurate and low-cost solutions to this problem, provided that one is willing to invest the effort required to develop sufficiently robust algorithms.
In this tutorial, we will therefore discuss some of the most promising approaches, their strengths, and their weaknesses.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Computer Vision]: Tracking

1. Introduction

Tracking an object in a video sequence means continuously identifying its location when either the object or the camera are moving. More specifically, 3D tracking aims at continuously recovering all six degrees of freedom that define the camera position and orientation relative to the scene, or, equivalently, the 3D displacement of an object relative to the camera.

Many other technologies besides vision have been tried to achieve this goal, but they all have their weaknesses: Mechanical trackers are accurate enough, although they tether the user to a limited working volume. Magnetic trackers are vulnerable to distortions by metal in the environment, which are a common occurrence, and also limit the range of displacements. Ultrasonic trackers suffer from noise and tend to be inaccurate at long ranges because of variations in the ambient temperature. Inertial trackers drift with time.

By contrast, vision has the potential to yield non-invasive, accurate and low-cost solutions to this problem, provided that one is willing to invest the effort required to develop sufficiently robust algorithms. In some cases, it is acceptable to add fiducials, such as LEDs or special markers, to the scene or target object to ease the registration task, as will be discussed in Section 2. Of course, this assumes that one or more fiducials are visible at all times. Otherwise, the registration falls apart. Moreover, it is not always possible to place fiducials. For example, Augmented Reality end-users do not like markers because they are visible in the scene and it is not always possible to modify the environment before the application has to run.

It is therefore much more desirable to rely on naturally present features, such as edges, corners, or texture. Of course, this makes tracking far more difficult: Finding and following feature points or edges on many every day's objects is sometimes difficult because there may be few of them. Total, or even partial occlusion of the tracked objects typically results in tracking failure. The camera can easily move too fast so that the images are motion blurred; the lighting during a shot can change significantly; reflections and specularities may confuse the tracker. Even more importantly, an object may drastically change its aspect very quickly due to displacement. For example this happens when a camera films a building and goes around the corner, causing one wall to disappear and a new one to appear. In such cases, the features to be followed always change and the tracker must deal with features coming in and out of the picture. Sections 3 and 4 focus on solutions to these difficult problems.

2. Fiducial-Based Tracking

Vision-based 3D tracking can be decomposed into two main steps: First image processing to extract some information from the images, and second pose estimation itself. The addition in the scene of *fiducials*, also called *landmarks* or *markers*, greatly helps both steps: They constitute image features easy to extract, and they provide reliable, easy to exploit measurements for pose estimation.

2.1. Point-Like Fiducials

Fiducials have been used for many years by close-range photogrammetrists. They can be designed in such a way that they can be easily detected and identified with an *ad hoc* method. Their image locations can also be measured to a much higher accuracy than natural features. In particular, circular fiducials work best, because the appearance of circular patterns is relatively invariant to perspective distortion, and because their centroid provides a stable 2D position, which can easily be determined with sub-pixel accuracy. The 3D positions of the fiducials in the world coordinate system are assumed to be precisely known: This can be achieved by hand, with a laser, or with a structure-from-motion algorithm. To facilitate their identification, the fiducials can be arranged in a distinctive geometric pattern. Once the fiducials are identified in the image, they provide a set of correspondences that can be used to retrieve the camera pose.

For high-end applications, companies such as Geodetic services, Inc., Advanced Real-time Tracking GmbH, Metronor, ViconPeak, AICON 3D Systems GmbH propose commercial products based on this approach. Lower-cost, and lower-accuracy solutions, have also been proposed by the Computer Vision community. For example, the Concentric Contrasting Circle (CCC) fiducial [HNL96] is formed by placing a black ring on a white background, or vice-versa. To detect these fiducials, the image is first thresholded, morphological operations are then applied to eliminate too small regions, and a connected component labeling operation is performed to find white and black regions, as well as their centroids. Along the same lines, [SHC*96] uses color-coded fiducials for a more reliable identification. Each fiducial consists of an inner dot and a surrounding outer ring, four different colors are used, and thus 12 unique fiducials can be created and identified based on their two colors. Because the tracking range is constrained by the detectability of fiducials in input images, [CLN98] introduces a system that uses several sizes for the fiducials. They are composed of several colored concentric rings, where large fiducials have more rings than smaller ones, and diameters of the rings are proportional to their distance to the fiducial center, to facilitate their identification. When the camera is close to fiducials, only small size fiducials are detected. When it is far from them, only large size fiducials are detected.

While all the previous methods for fiducial detection use *ad hoc* schemes, [CF04] uses a machine learning approach which delivers significant improvements in reliability. The fiducials are made of black disks on white background, and sample fiducial images are collected under varying perspective, scale and lighting conditions, as well as negative training images. A cascade of classifiers is then trained on these data: The first step is a fast Bayes decision rule classification, the second one a powerful but slower nearest neighbor classifier on the subset passed by the first stage. At run-time, all the possible sub-windows in the image are classified using this cascade. This results in a remarkably reliable fiducial detection method.

2.2. Extended Fiducials

The fiducials presented above were all circular and only their center was used. By contrast, [KKR*97] introduces squared, black on white, fiducials, which contain small red squares for their identification. The corners are found by fitting straight line segments to the maximum gradient points on the border of the fiducial. Each of the four corners of such fiducials provides one correspondence and the pose is estimated using an Extended Kalman filter.

[Rek98, KB99, KBP*00] also use planar, rectangular fiducials, and show that a single fiducial is enough to estimate the pose. Fig. 1 depicts their approach. It has become popular, because it yields a robust, low-cost solution for real-time 3D tracking, and a software library called ARToolKit is publicly available [ART].

The whole process, the detection of the fiducials and the pose estimation, runs in real-time, and therefore can be applied in every frame: The 3D tracking system does not require any initialization by hand, and is robust to fiducial occlusion. In practice, under good lighting conditions, the recovered pose is also accurate enough for Augmented Reality applications. These characteristics make ARToolKit a good solution to 3D tracking, whenever the engineering of the scene is possible.

3. Using Natural Features

Using markers to simplify the 3D tracking task requires engineering of the environment, which end-users of tracking technology do not like or is sometimes even impossible, for example in outdoor environments. Whenever possible, it is therefore much better to be able to rely on features naturally present in the images. Of course, this approach makes tracking much more challenging and some 3D knowledge is often

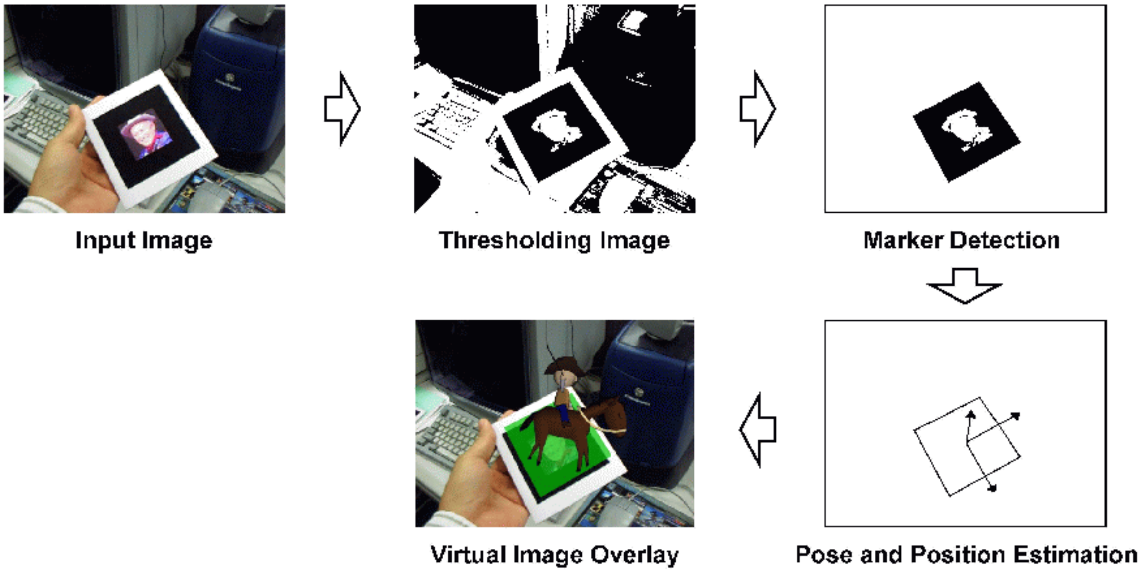


Figure 1: *Processing flow of ARToolKit: The marker is detected in the thresholded image, and then used to estimate the camera pose. (From [KBP*00], figure courtesy of H. Kato, M. Billinghurst, I. Poupyrev, K. Imamoto and K. Tachibana.)*

required to make things easier. For MR applications, this is not issue since 3D scene models are typically available and we therefore focus here on model-based approaches.

We distinguish here two families of approaches depending on the nature of the image features being used. The first one is formed by edge-based methods that match the projections of the target object 3D edges to area of high image gradient. The second family includes all the techniques that rely on information provided by pixels inside the object's projection.

3.1. Edge-Based Methods

Historically, the early approaches to tracking were all edge-based mostly because these methods are both computationally efficient, and relatively easy to implement. They are also naturally stable to lighting changes, even for specular materials, which is not necessarily true of methods that consider the internal pixels, as will be discussed later. The most popular approach is to look for strong gradients in the image around a first estimation of the object pose, without explicitly extracting the contours [Har92, AZ95, MBC01, DC02, CMC03, VLF04a], which is fast and general.

3.1.1. RAPiD

Even though RAPiD [Har92] was one of the first 3D tracker to successfully run in real-time and many improvements have been proposed since, many of its basic components have been retained in more recent systems. The key idea is to consider a set of 3D points on the object, called control points, which lie on high contrast edges in the images. As shown in Figure 2, the control points can be sampled along the 3D model edges and in the areas of rapid albedo change. They can also be generated on the fly as points on the occluding contours of the object. The 3D motion of the object between two consecutive frames can be recovered from the 2D displacement of the control points.

Once initialized, the system performs a simple loop: For each frame, the predicted pose, which can simply be the pose estimated for the previous frame, is used to predict which control points will be visible and what their new locations should be. The control points are matched to the image contours, and the new pose estimated from these correspondences via least-squares minimization.

In [Har92], some enhancements to this basic approach are proposed. When the edge response at a control point becomes too weak, it is not taken into account into the motion computation, as it may subsequently incorrectly latch on to a stronger nearby edge. As we will see below, this can also be handled using a robust estimator. An additional clue that can be used to reject incorrect edges is their polarity, that is whether they correspond to a transition from dark to light or from light to dark. A way to use occluding contours of the object is also given.

3.1.2. Making RAPiD Robust

The main drawback of of the original RAPiD formulation is its lack of robustness. The weak contours heuristics is not enough to prevent incorrectly detected edges from disturbing the pose computation. In practice, such errors are frequent.

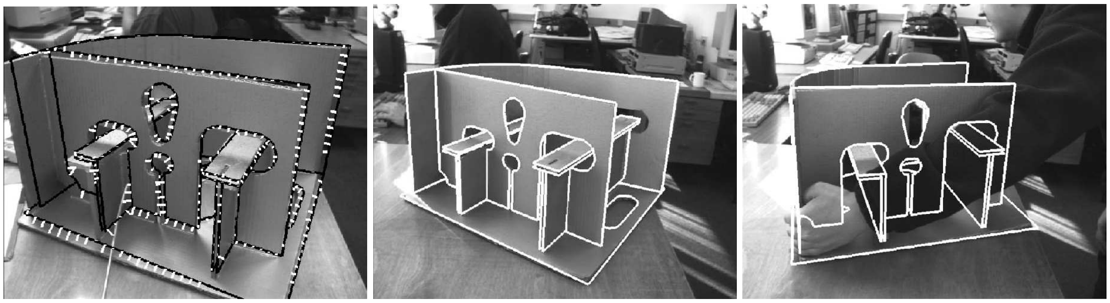


Figure 2: *In RAPiD-like approaches, control points are sampled along the model edges. The small white segments in the left image join the control points in the previous image to their found position in the new image. The pose can be inferred from these matches, even in presence of occlusions by introducing robust estimators. (From [DC02], figure courtesy of T. Drummond and R. Cipolla).*

They arise from occlusions, shadows, texture on the object itself, or background clutter.

Several methods have been proposed to make the RAPiD computation more robust. [DC02] uses a robust estimator and replaces the least-squares estimation by an iterative reweighted least-squares to solve the new problem. Similarly, [MBC01] uses a framework similar to RAPiD to estimate a 2D affine transformation between consecutive frames, but also replaces standard least-squares by robust estimation.

In the approaches described above, the control points were treated individually, without taking into account that several control points are often placed on the same edge, and hence their measurements are correlated. By contrast, in [AZ95, SB98] control points lying on the same object edge are grouped into primitives, and a whole primitive can be rejected from the pose estimation. In [AZ95], a RANSAC methodology [FB81] is used to detect outliers among the control points forming a primitive. If the number of remaining control points falls below a threshold after elimination of the outliers, the primitive is ignored in the pose update. Using RANSAC implies that the primitives have an analytic expression, and precludes tracking free-form curves. By contrast, [SB98] uses a robust estimator to compute a local residual for each primitive. The pose estimator then takes into account all the primitives using a robust estimation on the above residuals.

When the tracker finds multiple edges within its search range, it may end-up choosing the wrong one. To overcome this problem, in [DC02], the influence of a control point is inversely proportional to the number of edge strength maxima visible within the search path. [VLF04a] introduces another robust estimator to handle multiple hypotheses and retain all the maxima as possible correspondents in the pose estimation.

3.2. Texture-Based Methods

If the object is sufficiently textured, information can be can be derived from optical flow, template matching or interest-point correspondences. However the latter is probably the most effective for MR applications because they rely on matching local features. Given such correspondences, the pose can be estimated by least-square minimization, or even better, by robust estimation. They are therefore relatively insensitive to partial occlusions or matching errors. Illumination invariance is also simple to achieve. And, unlike edge-based methods, they do not get confused by background clutter and exploit more of the image information, which tends to make them more dependable.

3.3. Interest Point Detection and 2D Matching

In interest point methods, instead of matching all pixels in an image, only some pixels are first selected with an “interest operator” before matching. This reduces the computation time while increasing the reliability if the pixels are correctly chosen. [För86] presents the desired properties for such an interest operator: Selected points should be different from their neighbors, which eliminates edge-points; the selection should be repeatable, that is the same points should be selected in several images of the same scene, despite perspective distortion or image noise. In particular, the precision and the reliability of the matching directly depends on the invariance of the selected position. Pixels on repetitive patterns should also be rejected or at least given less importance to avoid confusion during matching.

Such an operator was already used in the 70’s for tracking purposes [Mor77, Mor81]. Numerous other methods have been proposed since and [DG93, SB95] give good surveys of them. Most of them involve second order derivatives, and results can be strongly affected by noise. Several successful interest point detectors [För86, HS88, ST94] rely on the auto-correlation matrix, whose coefficients are sums over a window of the first derivatives of the image intensity with respect to the pixel coordinates. The derivatives can be weighted using a Gaussian kernel to increase robustness to noise [SM97]. The derivatives should also be computed using a first order Gaussian kernel. This comes at a price since it can reduce the localization accuracy, and degrades the performance of the image patch correlation considered by the point matching procedure.

As discussed in [För86], the pixels can be classified from the behavior of the eigen values of the auto-correlation matrix: Pixels with two large, approximately equal eigen values are good candidates for selection. [ST94] shows that locations with two large eigenvalues can be reliably tracked, especially under affine deformations, and considers locations where the smallest eigen value is higher than a threshold. Interest points can then taken to be the locations that are local maxima of the chosen measure above a predefined threshold. It should be noted that these measures have a relatively poor localization accuracy and are computationally demanding. However they are widely used and certainly remain the best choice for tracking purposes.

For tracking purpose, it is then useful to match two sets of interest points and extracted from two images taken from similar viewpoints. A classical procedure [ZDFL95] runs as follows: For each point in the first image, search in a region of the second image around its location for a corresponding point. The search is based on the similarity of the local image windows centered on the points, which strongly characterise the points when the images are sufficiently close. The similarity can be measured using the zero-normalized cross-correlation that is invariant to affine changes of the local image intensities, and make the procedure robust to illumination changes. To obtain a more reliable set of matches, one can reverse the role of the two images, and repeat the previous procedure. Only the correspondences between points that chose each other are kept.

3.4. Eliminating Drift

In the absence of points whose coordinates are known *a priori*, all methods are subject to error accumulation, which eventually results in tracking failure and precludes of truly long sequences.

A solution to this problem is to introduce one or more *keyframes* such as the one in the upper left corner of Figure 3, that is images of the target object or scene for which the camera has been registered *beforehand*. At runtime, incoming images can be matched against the keyframes to provide a position estimate that is drift-free [RDLW95, GRSN02, TMdCM02]. This, however, is more difficult than matching against immediately preceding frames as the difference in viewpoint is likely to be much larger. The algorithm used to establish point correspondences must therefore both be fast and relatively insensitive to large perspective distortions, which is not usually the case for those used by the algorithms of Section 3.3 that need only handle small distortions between consecutive frames.

In [VLF04b], this is handled as follows. During a training stage, the system extracts interest points from each keyframe, back-projects them to the object surface to compute their 3D position, and stores image patches centered around their location. During tracking, for each new incoming image, the system picks the keyframe whose viewpoint is closest to that of the last known viewpoint. It synthesizes an *intermediate image* from that keyframe by warping the stored image patches to the last known viewpoint, which is typically the one corresponding to the previous image. The intermediate and the incoming images are now close enough that matching can be performed using simple, conventional, and fast correlation methods. Since the 3D position of keyframe interest has been precomputed, the pose can then be estimated by robustly minimizing the reprojection error. This approach handles perspective distortion, complex aspect changes, and self-occlusion. Furthermore, it is very efficient because it takes advantage of the large graphics capabilities of modern CPUs and GPUs.

However, as noticed by several authors [RDLW95, CCP02, TMdCM02, VLF04b], matching only against keyframes does not, by itself, yield directly exploitable results. This has two main causes. First, wide-baseline matching as described in the previous paragraph, is inherently less accurate than the short-baseline matching involved in frame-to-frame tracking, which is compounded by the fact that the number of correspondences that can be established is usually less. Second, if the pose is computed for each frame independently, no temporal consistency is enforced and the recovered motion can appear to be jerky. If it were used as is by an MR application, the virtual objects inserted in the scene would appear to *jitter*, or to tremble, as opposed to remaining solidly attached to the scene.

Temporal consistency can be enforced by some dynamical smoothing using a motion model. Another way proposed in [VLF04b] is to combine the information provided by the keyframes, which provides robustness, with that coming from preceding frames, which enforces temporal consistency. This does not make assumptions on the camera motion and improves the accuracy of the recovered pose. It is still compatible with the use of dynamical smoothing that can be useful to in case where the pose estimation remains unstable, for example when the object is essentailly fronto-parallel.

4. Tracking by Detection

The recursive nature of traditional 3D tracking approaches provides a strong prior on the pose for each new frame and makes image feature identifications relatively easy. However, it comes at a price: First, the system must either be initialized by hand or require the camera to be very close to a specified position. Second, it makes the system very fragile.

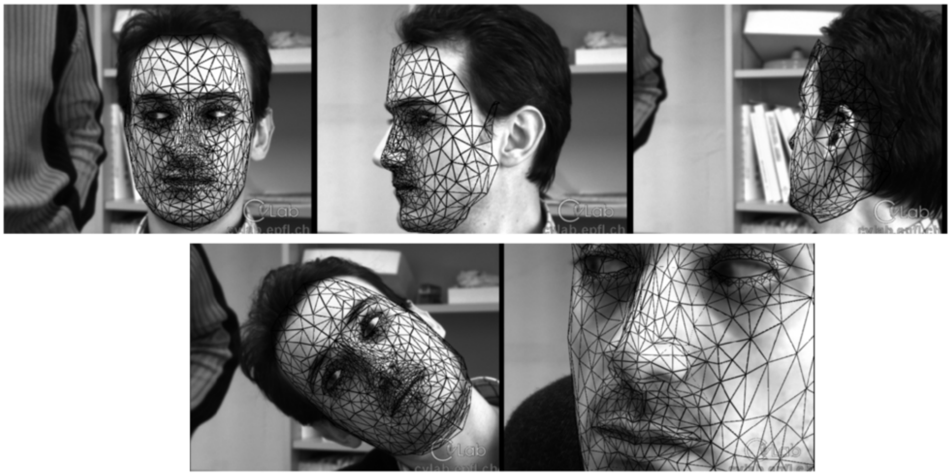


Figure 3: *Face tracking using interest points and one reference image shown on the top left. (From [VLF04b].)*

If something goes wrong between two consecutive frames, for example due to a complete occlusion of the target object or a very fast motion, the system can be lost and must be re initialized in the same fashion. In practice, such weaknesses make purely recursive systems nearly unusable, and the popularity of ARToolKit [KBP*00] in the Augmented Reality community should come as no surprise: It is the first vision-based system to really overcome these limitations by being able to detect the markers in every frame without constraints on the camera pose.

However, achieving the same level of performance *without* having to engineer the environment remains a desirable goal. Pose estimation from natural features without prior on the actual position is closely related to object detection and recognition. Object detection has a long history in Computer Vision, mostly focused on 2D detection even for 3D objects [NNM96, VJ01]. Nevertheless, there has been longstanding interest in simultaneous object detection and pose estimation. Early approaches were edge-based [Low91, Jur98], but methods based on feature points matching have become popular since [SM97] shows that local invariants work better than raw patches for such purpose. [SM97] uses invariants based on rotation invariant combination of image derivatives but other local invariants have been proposed. Considering feature point appear to be a better approach to achieve robustness to scale, viewpoint, illumination changes and partial occlusions than edge- or eigen-image- based techniques.

During an offline training stage, one builds a database of interest points lying on the object and whose position on the object surface can be computed. A few images in which the object has been manually registered are often used for this purpose. At runtime, feature points are first extracted from individual images and matched against the database. The object pose can then be estimated from such correspondences, for example using RANSAC [FB81] to eliminate spurious correspondences.

The difficulty in implementing such approaches comes from the fact that the database images and the input ones may have been acquired from very different viewpoints. As discussed in Section 3.3, unless the motion is very quick, this problem does not arise in conventional recursive tracking approaches because the images are close to each other. However, for tracking-by-detection purposes, the so-called *wide baseline* matching problem becomes a critical issue that must be addressed.

In the remainder of this section, we discuss in more detail the extraction and matching of feature points in this context. We conclude by discussing the relative merits of tracking-by-detection and recursive tracking.

4.1. Feature Point Extraction

To handle as wide as possible a range of viewing conditions, feature point extraction should be insensitive to scale, viewpoint, and illumination changes.

As proposed in [Lin94], scale-invariant extraction can be achieved by taking feature points to be local extrema of a Laplacian-of-Gaussian pyramid in scale-space. To increase computational efficiency, the Laplacian can be approximated by a Difference-of-Gaussians [Low99]. Research has then focused on affine invariant region detection to handle more perspective changes. [Bau00, SZ02, MS02] used an affine invariant point detector based on the Harris detector, where the affine transformation that makes equal the two eigen values of the auto correlation matrix is evaluated to rectify the patch appearance. [TV00] achieves such invariance by fitting an ellipse to the local texture. [MCMP02] proposes a fast algorithm to extract Maximally Stable Extremal Regions demonstrated in a live demo. [MTS*05] gives a good summary and comparisons of the existing affine invariant regions detectors.

4.2. Wide Baseline Matching

Once a feature point has been extracted, the most popular approach to matching it is first to characterize it in terms of its image neighborhood and then to compare this characterization to those present in the database. Such characterization, or *local descriptor*, should be not only invariant to viewpoint and illumination changes but also highly distinctive. We briefly review some of the most representative below.

4.2.1. Local Descriptors

Many such descriptors have been proposed over the years. For example, [SM97] computes rotation invariant descriptors as functions of relatively high order image derivatives to achieve orientation invariance; [TV00] fits an ellipse to the texture around local intensity extrema and uses the Generalized Color Moments [MMV99] as a descriptor. [Low04] introduces a descriptor called SIFT based on multiple orientation histograms, which tolerates significant local deformations. This last descriptor has been shown in [MS03] to be one of the most efficient. As illustrated by Figure 4, it has been successfully applied to 3D tracking in [SLL02, SL04] and we now describe it in more detail.

The remarkable invariance of the SIFT descriptor is achieved by a succession of carefully designed techniques. First the location and scale of the keypoints are determined precisely by interpolating the pyramid of Difference-of-Gaussians used for the detection. To achieve image rotation invariance, an orientation is also assigned to the keypoint. It is taken to be the one corresponding to a peak in the histogram of the gradient orientations within a region around the keypoint. This method is quite stable under viewpoint changes, and achieves an accuracy of a few degrees. The image neighborhood of the feature point is then corrected according to the estimated scale and orientation, and a local descriptor is computed on the resulting image region to achieve invariance to the remaining variations, such as illumination or out-of-plane variation. The point neighborhood is divided into several, typically 4×4, subregions and the contents of each subregion is summarized by an height-bin histogram of gradient orientations. The keypoint descriptor becomes a vector with 128 dimensions, built by concatenating the different histograms. Finally, this vector is normalized to unit length to reduce the effects of illumination changes.

4.2.2. Statistical Classification

The SIFT descriptor has been empirically shown to be both very distinctive and computationally cheaper than those based on filter banks. To shift even more of the computational burden from matching to training, which can be performed beforehand, we have proposed in our own work an alternative approach based on machine learning techniques [LLF05]. We treat wide baseline matching of keypoints as a classification problem, in which each class corresponds to the set of all possible views of such a point. Given one or more images of a target object, the system synthesizes a large number of views, or image patches, of individual keypoints to automatically build the training set. If the object can be assumed to be locally planar, this is done by simply warping image patches around the points under affine deformations, otherwise, given the 3D model, standard Computer Graphics texture-mapping techniques can be used. This second approach relaxes the planarity assumptions.

The classification itself is performed using randomized trees [AG97]. Each non-terminal node of a tree contains a test of the type: “Is this pixel brighter than this one ?” that splits the image space. Each leaf contains an estimate based on training data of the conditional distribution over the classes given that a patch reaches that leaf. A new image is classified by simply dropping it down the tree. Since only pixel intensities comparisons are involved, this procedure is very fast and robust to illumination changes. Thanks to the efficiency of randomized trees, it yields reliable classification results.

As depicted by Figure 5, this method has been successfully used to detect and compute the 3D pose of planar, non-planar, and even deformable objects [LLF05, PLF05].

4.2.3. From Wide Baseline Matching to 3D Tracking

As mentioned before, wide baseline matching techniques can be used to perform 3D tracking. To illustrate this, we briefly describe here the SIFT-based implementation reported in [SL04].

First, during a learning stage, a database of scene feature points is built by extracting SIFT keypoints in some reference images. Because the keypoints are detected in scale-space, the scene does not necessarily have to be well-textured. Their 3D positions are recovered using a structure-from-motion algorithm. Two-view correspondences are first established based on the SIFT descriptors, and chained to construct multi-view correspondences while avoiding prohibitive complexity. Then the 3D positions are recovered by a global optimization over all camera parameters and these point coordinates, which is initialized as suggested in [SK94]. At run-time, SIFT features are extracted from the current frame, matched against the database, resulting in a set of 2D / 3D correspondences that can be used to recover the pose.

The best candidate match for a SIFT feature extracted from the current frame is assumed to be its nearest neighbor, in the sense of the Euclidean distance of the descriptor

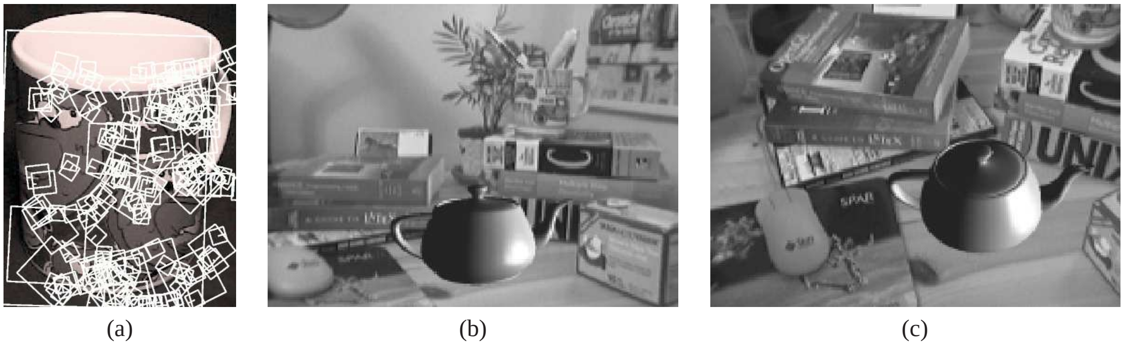


Figure 4: *Using SIFT for tracking-by-detection. (a) Detected SIFT features [Low01]. (b,c) They have been use to track the pose of the camera and add the virtual teapot [SL04]. (Courtesy of D.G. Lowe and I. Gordon).*

vectors, in the point database. The size of the database and the high dimensionality of these vectors would make the exhaustive search intractable, especially for real-time applications. To allow for fast search, the database is organized as a k-d tree. The search is performed so that bins are explored in the order of their closest distance from the query description vector, and stopped after a given number of data points has been considered, as described in [BL97]. In practice, this approach returns the actual nearest neighbor with high probability.

As discussed Section 3.4, recovering the camera positions in each frame independently and from noisy data typically results in jitter. To stabilize the pose, a regularization term that smoothes camera motion across consecutive frames is introduced. Its weight is iteratively estimated to eliminate as much jitter as possible without introducing drift when the motion is fast. The full method runs at four frames per second on a 1.8 GHz ThinkPad.

4.3. The End of Recursive Tracking?

Since real-time tracking-by-detection has become a practical possibility, one must wonder if the conventional recursive tracking methods that have been presented in the previous sections of this survey are obsolescent.

We do not believe this to be the case. As illustrated by the case of the SIFT-based tracking system [SL04] discussed above, treating each frame independently has its problems. Imposing temporal continuity constraints across frames can help increase the robustness and quality of the results. Furthermore, wide baseline matching tends to be both less accurate and more computationally intensive than the short baseline variety.

As shown in Section 3.4, combining both kinds of approaches can yield the best of both worlds: Robustness from tracking-by-detection, and accuracy from recursive tracking. In our opinion, this is where the future of tracking lays. The challenge will be to become able, perhaps by taking advantage of recursive techniques that do not require prior training, to learn object descriptions online so that a tracker can operate in a complex environment with minimal *a priori* knowledge.

5. Conclusion

Even after more than twenty years of research, practical vision-based 3–D tracking systems still rely on fiducials because this remains the only approach that is sufficiently fast, robust, and accurate. Therefore, if it is practical to introduce them in the environment the system inhabits, this solution surely must be retained. ARToolkit [ART] is a freely available alternative that uses planar fiducials that may be printed on pieces of paper. While less accurate, it remains robust and allows for fast development of low-cost applications. As a result, it has become popular in the Augmented Reality Community.

However, this state of affairs may be about to change as computers have just now become fast enough to reliably handle natural features in real-time, thereby making it possible to completely do away with fiducials. This is especially true when dealing with objects that are polygonal, textured, or both [DC02, VLF04b]. However, the reader must be aware that the recursive nature of most of these algorithms makes them inherently fragile: They must be initialized manually and cannot recover if the process fails for any reason. In practice, even the best methods suffer such failures all too often, for example because the motion is too fast, a complete occlusion occurs, or simply because the target object moves momentarily out of the field of view.

This can be addressed by combining image data with dynamics data provided by inertial sensors or gyroscopes [SHC*96, FN03]. The sensors allow a prediction of the camera position or relative motion that can then be refined using vision techniques similar to the ones described in this survey. Such combination is possible for applications,

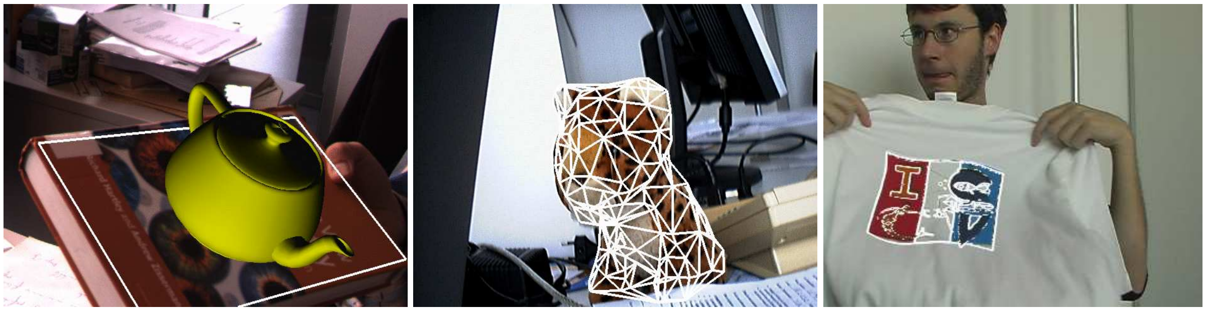


Figure 5: *Detection and computation in real-time of the 3D pose of a planar object, a full 3D object, and a deformable object. (From [LLF05] and [PLF05].)*

such as Augmented Reality, that require tracking of the camera with respect to a static scene, assuming one is willing to instrument the camera. However, instrumenting the camera is not always an option. For example, it would be of no use to track moving cars with a static camera.

A more generic and desirable approach is therefore to develop purely image-based methods that can detect the target object and compute its 3D pose from a single image. If they are fast enough, they can then be used to initialize and re-initialize the system as often as needed, even if they cannot provide the same accuracy as traditional recursive approaches that use temporal continuity constraints to refine their estimates. Techniques able to do just this are just beginning to come online [SL04, LLF05]. And, since they are the last missing part of the puzzle, we expect that we will not have to wait for another twenty years for purely vision-based commercial systems to become a reality.

References

[AG97] Amit Y., Geman D.: Shape quantization and recognition with randomized trees. *Neural Computation 9*, 7 (1997), 1545–1588. 7

[ART] Artoolkit. http://www.hitl.washington.edu/artoolkit/. 2, 8

[AZ95] Armstrong M., Zisserman A.: Robust object tracking. In *Proceedings of Asian Conference on Computer Vision* (1995), pp. 58–62. 3, 4

[Bau00] Baumberg A.: Reliable feature matching across widely separated views. In *Conference on Computer Vision and Pattern Recognition* (2000), pp. 774–781. 6

[BL97] Beis J., Lowe D.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Conference on Computer Vision and Pattern Recognition* (Puerto Rico, 1997), pp. 1000–1006. 8

[CCP02] Chia K., Cheok A., Prince S.: Online 6 dof augmented reality registration from natural features. In *International Symposium on Mixed and Augmented Reality* (2002). 5

[CF04] Claus D., Fitzgibbon A.: Reliable fiducial detection in natural scenes. In *European Conference on Computer Vision* (May 2004), vol. 3024, Springer-Verlag, pp. 469–480. 2

[CLN98] Cho Y., Lee W., Neumann U.: A multi-ring color fiducial system and intensity-invariant detection method for scalable fiducial-tracking augmented reality. In *International Workshop on Augmented Reality* (1998). 2

[CMC03] Comport A., Marchand E., Chaumette F.: A real-time tracker for markerless augmented reality. In *International Symposium on Mixed and Augmented Reality* (Tokyo, Japan, September 2003). 3

[DC02] Drummond T., Cipolla R.: Real-time visual tracking of complex structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 7 (july 2002), 932–946. 3, 4, 8

[DG93] Deriche R., Giraudon G.: A computational approach for corner and vertex detection. *International Journal of Computer Vision 10*, 2 (1993), 101–124. 4

[FB81] Fischler M., Bolles R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications ACM 24*, 6 (1981), 381–395. 4, 6

[FN03] Foxlin E., Naimark L.: Miniaturization, claibration and accuracy evaluation of a hybrid self-tracker. In *International Symposium on Mixed and Augmented Reality* (Tokyo, Japan, 2003). 8

[För86] FÖRSTNER W.: A feature-based correspondence algorithm for image matching. *International Archives of Photogrammetry and Remote Sensing 26*, 3 (1986), 150–166. 4, 5

[GRSN02] GENC Y., RIEDEL S., SOUVANNAVONG F., NAVAB N.: Marker-less tracking for augmented reality: A learning-based approach. In *International Symposium on Mixed and Augmented Reality* (2002). 5

[Har92] HARRIS C.: *Tracking with Rigid Objects*. MIT Press, 1992. 3

[HNL96] HOFF W. A., NGUYEN K., LYON T.: Computer vision-based registration techniques for augmented reality. In *Proceedings of Intelligent Robots and Control Systems XV, Intelligent Control Systems and Advanced Manufacturing* (November 1996), pp. 538–548. 2

[HS88] HARRIS C., STEPHENS M.: A combined corner and edge detector. In *Fourth Alvey Vision Conference, Manchester* (1988). 4

[Jur98] JURIE F.: Tracking objects with a recognition algorithm. *Pattern Recognition Letters 3-4*, 19 (1998), 331–340. 6

[KB99] KATO H., BILLINGHURST M.: Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *IEEE and ACM International Workshop on Augmented Reality* (October 1999). 2

[KBP*00] KATO H., BILLINGHURST M., POUPYREV I., IMAMOTO K., TACHIBANA K.: Virtual object manipulation on a table-top ar environment. In *International Symposium on Augmented Reality* (2000), pp. 111–119. 2, 3, 6

[KKR*97] KOLLER D., KLINKER G., ROSE E., BREEN D., WHITAKER R., TUCERYAN M.: Real-time vision-based camera tracking for augmented reality applications. In *ACM Symposium on Virtual Reality Software and Technology* (Lausanne, Switzerland, September 1997), pp. 87–94. 2

[Lin94] LINDEBERG T.: Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics 21*, 2 (1994), 224–270. 6

[LLF05] LEPETIT V., LAGGER P., FUA P.: Randomized trees for real-time keypoint recognition. In *Conference on Computer Vision and Pattern Recognition* (San Diego, CA, June 2005). 7, 9

[Low91] LOWE D. G.: Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*, 5 (June 1991), 441–450. 6

[Low99] LOWE D.: Object recognition from local scale-invariant features. In *International Conference on Computer Vision* (1999), pp. 1150–1157. 6

[Low01] LOWE D.: Local feature view clustering for 3d object recognition. In *Conference on Computer Vision and Pattern Recognition* (2001), vol. 1, pp. 682–688. 8

[Low04] LOWE D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 20*, 2 (2004), 91–110. 7

[MBC01] MARCHAND E., BOUTHEMY P., CHAUMETTE F.: A 2d-3d model-based approach to real-time visual tracking. *Image and Vision Computing 19*, 13 (2001), 941–955. 3, 4

[MCMP02] MATAS J., CHUM O., MARTIN U., PAJDLA T.: Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference* (London, UK, September 2002), pp. 384–393. 6

[MMV99] MINDRU F., MOONS T., VANGOOL L.: Recognizing color patterns irrespective of viewpoint and illumination. In *Conference on Computer Vision and Pattern Recognition* (1999), pp. 368–373. 7

[Mor77] MORAVEC H.: Towards automatic visual obstacle avoidance. In *International Joint Conference on Artificial Intelligence* (MIT, Cambridge, Mass., August 1977), p. 584. 4

[Mor81] MORAVEC H.: *Robot Rover Visual Navigation*. UMI Research Press, Ann Arbor, Michigan, 1981. 4

[MS02] MIKOLAJCZYK K., SCHMID C.: An affine invariant interest point detector. In *European Conference on Computer Vision* (2002), Springer, pp. 128–142. Copenhagen. 6

[MS03] MIKOLAJCZIK K., SCHMID C.: A performance evaluation of local descriptors. In *Conference on Computer Vision and Pattern Recognition* (June 2003), pp. 257–263. 7

[MTS*05] MIKOLAJCZYK K., TUYTELAARS T., SCHMID C., ZISSERMAN A., MATAS J., SCHAFFALITZKY F., KADIR T., GOOL L. V.: A comparison of affine region detectors. *Accepted to International Journal of Computer Vision* (2005). 7

[NNM96] NAYAR S. K., NENE S. A., MURASE H.: Real-time 100 object recognition system. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*, 12 (1996), 1186–1198. 6

[PLF05] PILET J., LEPETIT V., FUA P.: Real-time non-rigid surface detection. In *Conference on Computer Vision and Pattern Recognition* (San Diego, CA, June 2005). 7, 9

[RDLW95] RAVELA S., DRAPER B., LIM J., WEISS R.: Adaptive tracking and model registration across distinct aspects. In *International Conference on Intelligent Robots and Systems* (1995), pp. 174–180. 5

[Rek98] REKIMOTO J.: Matrix: A realtime object identification and registration method for augmented reality. In *Asia Pacific Computer Human Interaction* (1998). 2

[SB95] SMITH S. M., BRADY J. M.: *SUSAN – A new approach to low level image processing*. Tech. Rep. TR95SMS1c, Oxford University, Chertsey, Surrey, UK, 1995. 4

[SB98] SIMON G., BERGER M.-O.: A two-stage robust statistical method for temporal registration from features of various type. In *International Conference on Computer Vision* (Bombay, India, Jan. 1998), pp. 261–266. 4

[SHC*96] STATE A., HIROTA G., CHEN D., GARETT W., LIVINGSTON M.: Superior augmented reality registration by integrating landmark tracking and magnetic tracking. *Computer Graphics, SIGGRAPH Proceedings* (July 1996), 429–438. 2, 8

[SK94] SZELISKI R., KANG S.: Recovering 3–d shape and motion from image streams using non linear least squares. *Journal of Visual Communication and Image Representation 5*, 1 (1994), 10–28. 7

[SL04] SKRYPNYK I., LOWE D. G.: Scene modelling, recognition and tracking with invariant image features. In *International Symposium on Mixed and Augmented Reality* (Arlington, VA, Nov. 2004), pp. 110–119. 7, 8, 9

[SLL02] SE S., LOWE D. G., LITTLE J.: Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research 22*, 8 (2002), 735–758. 7

[SM97] SCHMID C., MOHR R.: Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 5 (May 1997), 530–534. 5, 6, 7

[ST94] SHI J., TOMASI C.: Good features to track. In *Conference on Computer Vision and Pattern Recognition* (Seattle, June 1994). 4, 5

[SZ02] SCHAFFALITZKY F., ZISSERMAN A.: Multi-View Matching for Unordered Image Sets, or "How Do I Organize My holiday Snaps?". In *Proceedings of European Conference on Computer Vision* (2002), pp. 414–431. 6

[TMdCM02] TORDOFF B., MAYOL W., DE CAMPOS T., MURRAY D.: Head pose estimation for wearable robot control. In *British Machine Vision Conference* (2002), pp. 807–816. 5

[TV00] TUYTELAARS T., VANGOOL L.: Wide baseline stereo matching based on local, affinely invariant regions. In *British Machine Vision Conference* (2000), pp. 412–422. 6, 7

[VJ01] VIOLA P., JONES M.: Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition* (2001), pp. 511–518. 6

[VLF04a] VACCHETTI L., LEPETIT V., FUA P.: Combining edge and texture information for real-time accurate 3d camera tracking. In *International Symposium on Mixed and Augmented Reality* (Arlington, VA, November 2004). 3, 4

[VLF04b] VACCHETTI L., LEPETIT V., FUA P.: Stable real-time 3d tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 10 (October 2004), 1385–1391. 5, 6, 8

[ZDFL95] ZHANG Z., DERICHE R., FAUGERAS O., LUONG Q.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence 78* (1995), 87–119. 5

II-3: Perception and Sensors for Virtual Humans

Daniel Thalmann, Toni Condé

Virtual Reality Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)

Abstract
A Virtual Human is situated in a Virtual Environment (VE) equipped with sensors for vision, audition and tactile, informing it of the external VE and its internal state, it may also be aware of the user through real sensors like cameras and microphones. A Virtual Human possesses effectors, which allow it to exert an influence on the VE and a control architecture, which coordinates its perceptions and actions. In order to select the appropriate actions of an actor, the behavioral module needs to know the state of the environment of the actor. However, an actor is not passive, but performs actions which might involve objects, other actors, or even the user. Moreover, the actions of an actor may cause some events. Therefore, perception is decomposed into three categories: perception of objects and actors, actions of actors and events.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Three-Dimensional Graphics and Realism]: Virtual Humans, Mixed Reality, Sensors, Perception

1. Introduction

Behavioural Animation requires various capabilities, that we categorise into four elements: the awareness of the Autonomous Virtual Human (AVH) agent starts with a simulated perception of its environment (and the capability to memorise it). Based on this input, the agent will adapt its behaviour, take the proper decisions and then interact, either with elements of the environment or with other virtual humans (like when conversing, for example). Of course, it is not that easy to generate proper and convincing results without taking care of some animation problems, which are closer to Computer Graphics than AI.

We generally consider autonomy as the quality or state of being self-governing. As we said, it relies on different factors: perception of the elements in the environment is essential, as it gives the agent the awareness of what is changing around it. It is indeed the most important element that one should simulate before going further. Most common perceptions include (but are not limited to) simulated visual and auditive feedback. Adaptation and intelligence then define how the agent is capable of reasoning about what it perceives, especially when unpredictable events happen. On the other hand, when predictable elements are showing up again, it is necessary to have a memory capability, so that similar behaviour can be selected again. Lastly, emotion instantaneously adds realism by defining affective relationships between agents.

An AVH situated in a Virtual Environment (VE) is equipped with sensors for vision, audition and touch that inform it of the external VE and its internal state. An AVH possesses effectors to let it exert an influence on the VE and control architecture to coordinate its perceptions and actions. The behaviour of an AVH is adaptive as long as the control architecture allows it to maintain its variables in their validity zone.

The “mental processes“ of an AVH can be simulated. *Behavioural animation* includes the techniques applied to make an AVH intelligent and autonomous, to react to its VE and to make decisions based on its perceptive system, its short-term memory and long-term reasoning. *Intelligence* is the ability to plan and carry out the tasks based on the model of the current state of the VE.

Our objective is to permit the AVH to explore unknown VEs and to construct *mental structures and models, cognitive maps or plans from this exploration.* Once its representation has been created, the knowledge can be communicated to other AVHs. Each AVH perceives objects and other AVHs with the help of its VE, which provides information concerning their nature and positions. The behavioural model decides which action the AVH should take (such as walking or handling an object) and then uses the knowledge.

2. Virtual Sensors Background

An AVH should be equipped with virtual sensors for vision, audition and touch. These sensors constitute a starting point to implement behaviour such as direct vision during a move, handling of objects and responding to sounds or words.

After acquiring the information, the basic perceptive part of the AVH can be carried out by the *Flexible Perception Pipeline* approach proposed by [BBT99].

The perception of the AVH in its VE is communicated by vision and sound, sometimes by sensorial tactile information. Its behaviour, as in humans, is strongly influenced by data supplied by its sensors and its own intelligence for certain ends such as: extraction, simplification and filtering, which depend on perception criteria associated with each sensorial modality.

The AVH explores an unknown environment constructed on mental models as well as a "cognitive map" based on this exploration. Navigation is carried out in two ways: globally (with the pre-learned model of the VE, a few changes and the search for performance with a *path planning* algorithm) and locally (with direct acquisition of the VE). A 3D geometrical model in the form of a grid is implemented with the help of an *octree* combined with the approach proposed by [NRTM95][KT99].

2.1 Synthetic Vision

It is tempting to simulate perception by directly retrieving the location of each perceived object straight from the environment. This is of course the fastest solution (and has been extensively used in video-games until the mid-nineties) but no one can ever pretend that it is realistic at all (although it can be useful, as we will see later on). Consequently, various ways of simulating visual perception have been proposed, depending on whether geometric or semantic information (or both) are considered. We are going to compare now synthetic vision, geometric vision and database access, as described by Monzani [MGD03].

2.1.1 Synthetic vision through off-screen rendering

Synthetic vision, introduced by Renault et al. [RMT90], is achieved by rendering off-screen the scene as viewed by the agent. During the process, each individual object in the scene is assigned a different colour, so that once the 2D image has been computed, objects can still be identified: it is then easy to know which object is in sight by maintaining a table of correspondences between colours and objects' IDs. Furthermore, highly detailed depth information is retrieved from the view z-buffer, giving a precise location for each object.

Synthetic vision has been then successfully applied to animals (virtual fishes [TT94], and the SILAS dog [BTM96]). Noser et al. [NRTM95] showed how one can simulate a memory for agents: the 2D rendering and the corresponding z-buffer data are combined in order to determine whether the corresponding voxel of the scene is occupied by an object or not. By navigating through the environment, the agent will progressively construct a voxel-based representation of it. Of course, a rough implementation of this method would suffer from dramatic memory cost, because of the high volume required to store all voxels. Noser et al. proposed to use octrees instead which successfully reduces the amount of data. Once enough information has been gathered through exploration, the agent is then able to locate things and find its way. Another application of synthetic vision is real-time collision avoidance for multiple agents: in this case, each agent is perceiving the others, and dynamically creates local goals so that it avoids others while trying to reach its original global goal.

Synthetic vision is the most elegant method, because it is the more realistic simulation of vision and addresses correctly vision issues such as occlusion for instance. However, rendering the whole scene for each agent is very costly and for real-time applications, one tend to favour geometric vision.

2.1.2 Geometric vision

Bordeux et al. [BBT99], has proposed a perception pipeline architecture into which filters can be combined to extract the required information. The perception filter represents the basic entity of the perception mechanism. Such a filter receives a perceptible entity from the scene as input, extracts specific information about it, and finally decides to let it pass through or not. The criteria used in the decision process depends on the perception requirements. For virtual objects, they usually involve considerations about the distance and the relative direction of the object, but can also be based on shape, size, colour, or generic semantic aspects, and more generally on whatever the agent might need to distinguish objects. Filters are built with an object oriented approach: the very basic filter for virtual objects only considers the distance to the object, and its descendants refine further the selection.

Actually, the structure transmitted to a filter contains, along with the object to perceive, a reference to the agent itself and previously computed data about the object. The filter can extend the structure with the results of its own computation, for example the relative position and speed of the object, a probable time to impact or the angular extension of the object from the agent s point of view. Since a perception filter does not store data concerning the objects that passed through it, it is fully reentrant and can be used by several agents at the same time. This allows the creation of a common pool of filters at the application, each agent then referencing the filters it needs, thus avoiding useless duplication.

As an example of filters, Bordeux has implemented a basic range filter which selects objects in a given range around the agent. The field of view filter simulates an agent field of view with a given angular aperture. The collision filter detects potential impacts with other objects in the agent neighborhood and estimates, if needed, the time to impact, the object's relative speed and a local area to escape from. This has been used again in a safe-navigation behaviour which dynamically computes a collision-free path through the world. It is even possible to specify how long an object shall stay in the list after it was perceived, in order to simulate short-term memory.

However, the major problem with Geometric vision is to find the proper formulas when intersecting volumes (for instance, intersecting the view frustum of the agent with a volume in the scene). One can use bounding boxes to reduce the computation time, but it will always be less accurate than Synthetic vision. Nevertheless, it can be sufficient for many applications and, as opposed to Synthetic vision, the computation time can be adjusted precisely by refining the bounding volumes of objects.

2.1.3 Database access

Data access makes maximum use of the scene data available in the application, which can be distributed in several modules. For instance, the objects position, dimensions and shape are maintained by the rendering engine whereas semantic data about objects can be maintained by a completely separate part of the application. Due to scalability constraints as well as plausibility considerations, the agents generally restrain their perception to a local area around them instead of the whole scene. This method is generally chosen when the number of agents is high, like in Reynolds flocks of birds [Rey87], and schools of fishes. In [MT01] crowd simulation, agents directly know the position of their neighbours and compute coherent collision avoidance trajectory. As said before, the main problem with the method is the lack of realism, which can only be avoided by using one of the other methods.

2.2 Synthetic Audition

In real life, the behaviour of people or animals can be strongly influenced by sounds. Wenzel [Wen92] this "the function of the ears is to point the eyes". Audition is a temporal sensor, which is very sensitive to changes in acoustic signals. We can locate objects in space and, even more specifically, when they move. Moreover, acoustic signals carry a lot of semantic and emotional information; they inform us about sound sources relative to us as well as the propagation of sound paths in an acoustic environment. The restitution of sound must be very effective to react to sound events perceived by the AVHs in each frame.

The most important properties of a sound source in terms of computer knowledge are: 3D position of the source in the world, orientation of the sound source, cone of propagation, distance between the listener and the source, Doppler effect, volume and frequency, occlusion, obstruction and exclusion.

All these parameters can be set to filter the sound source depending on the simulation conditions. Regarding 3D sound, one may believe that it is sufficient to place the sound source in a 3D world without taking care of its direction [Car02]. However this is too big a limitation, especially for reverberations and reflections. In general, we represent the sound propagation with a cone. This solution gives us the flexibility to set specifically the different filters for each sound source.

2.3 Synthetic Tactile

Sensorial tactile information can be used to push buttons or to touch and handle objects. The simulation of this kind of sensor resembles the collision detection proposed by [HBMT95]. We can opt for the process described by [HLC*97] with V-Collide collision detection approach.

The V-Collide approach performs efficient and exact collision between triangulated polygonal models. It uses a 2-level hierarchical approach:

- The top level eliminates from consideration pairs of objects that are not close to each other,
- The bottom level performs exact collision detection down to the level of the triangles themselves.

2.4 Synthetic World

In synthetic vision, the vision models for an AVH are different from those used in behavioural robotics. A robot can only acquire information from its environment through these sensors, which limits its behaviour in navigating and avoiding obstacles. In a VE, supplementary information can be extracted and dealt with according to the perception model chosen for the AVH. This makes it faster and "intelligent" during actions. To optimise the model we chose a certain type of representation of the virtual world where the AVH maintains for example its vision at a low level system.

Nevertheless, because of scale and plausibility constraints of its autonomous behaviour, an AVH restricts its perception locally in relation to the VE as a whole. This approach is used more specifically when there are a lot of AVHs. Most important, the choice of the method must reflect an AVH in a VE close to reality.

3. Perception

As mentioned above, considerable restrictions appear when the actions produced by the AVH require dynamic knowledge of the VE with a perception system. One of these restrictions concerns the reflex actions, which require perception, but not memory concerning what has been perceived. In the classical approaches different behaviours implement their own perception mechanisms.

Perception of events is slightly more complex because events themselves are decomposed into three classes: desirable events, events happening to another actor and potential events which may or may not occur. The perception of the nature and the characteristics of an object, an actor or an action is not easily done from their 3D representation. Recognizing an action through motion is difficult as well. The adopted solution is to categorize every object, actor and action based on its nature and characteristics. We will then study the problem in the context of a group and even a large crowd.

Several methods like the one used by [BBT99] have been proposed to implement perception. The AVH maintains a perception puzzle, each piece corresponding to a specific *virtual sensor*. A pipeline is composed of filters to extract relevant information from the data supplied by the related *sensor*. An attempt to model an approach of unified perception is described by [CT04].

3.1 Proprioception

Proprioception is inspired by the human immune system composed of functional layers combining rapid and archaic mechanisms of innate immunity and slower mechanisms of acquired or adaptive immunity. The response time depends on the anterior exposition to pathogen.

Proprioception is based mainly on the integration in a same model of endogenous variables, homeostasis and reinforcement learning as proposed by [Ber94]:

- The notion of “endogenous“ variables captures information related to the internal state of the AVH that is influenced by both the AVH’s perceptions and actions but not restricted to them. Additional influences permit differentiation between the effects of similar perception inputs, on the resulting AVH's action. The existence of these variables constitutes an important type of cognitive intermediary between the sensory and virtual human controller poles of the behaviour loops.
- The notion of “homeostasis“ describes variables whose temporal dynamics must guarantee their keeping within pre-determined boundaries. Since exceeding these boundaries in human beings would result either in significant discomfort or in the death of the AVH, actions are taken to prevent these variables from departing from the set-value.

In order to be truly autonomous, the AVH must not just be capable of intelligent action, but also be self-sustaining. The third aspect of the model is the use of a reinforcement learning mechanism. This enables the discovery of a sequence of actions, which allows the AVH to remain viable despite the strong constraints exerted by the VE and the AVH's own endogenous variability.

3.2 Active Perception

In a VE, an AVH requires a combination of perception and action to behave in an autonomous way. The perception system provides a uniform interface to various techniques in the field of virtual perception, including synthetic vision, audition and touch. In usual approaches, different behaviours implement their own perception mechanisms, which leads to computation duplication when multiple behaviours are involved. Basically, an AVH maintains a set of perception pipelines, each corresponding to a particular type of virtual sensor [BBT99]. A pipeline is composed of filters that coordinate themselves in order to extract relevant information from the data sent by the associated virtual sensor.

3.3 Predictive Perception

The faculty of predicting, one of the main activities of the human brain is an essential notion in the perception of an AVH. It plays a basic role in active perception by giving the AVH the possibility to direct it’s look and attention elsewhere.

This prediction can be found in humans when anticipating the path a ball they should catch will take, avoiding mobile obstacles, preparing the body to wake up during the final hours of sleep, or even in the absurd effectiveness of a placebo [Lin01].

To obtain an active perception like directing the look and attention elsewhere, prediction functions must be organised. In the visual system alone, certain zones of the cortex deal with outlines, others with forms, movement, distance or colour. These processes are unconscious.

Predictive perception may be modeled using the mathematical theory of the observer. Algorithms are used to predict from partial measures, often external and with sound effects, the internal state of a non-linear system. An observer is typically composed of a system simulation that uses an internal model that may be approximate. It is guided and corrected by the measures taken on by the system. In problems of active perception and in certain circumstances, the observer also allows the selection of the measure or combination of measures to be carried out. This is particularly useful in improving the estimation of the system state at a given moment; this is inspired by the human nervous system.

4. Methodology

An AVH is not only situated in an environment with the help of virtual sensors but it must be equipped with behaviour, perception and memorisation faculties to make it autonomous and “intelligent“. Our objective is to give an AVH the ability to explore an unknown environment and thus construct mental models and “cognitive maps“. During or after the construction of these models, the AVH can carry out many functions successfully, for example “path-planning“, navigating and “place-finding“.

Our model is based on the multi-sensory integration of the standard theory of neuroscience (Figure 1). Signals related to the same virtual object but coming from distinct sensory systems are combined. We will focus on two aspects: 1) the assignment problem: determining which sensory stimuli belong to the same virtual object and 2) the sensory recoding problem: recoding signals in a common format before combining them [Pou02].

We can consider a multi-sensorial approach based on a 3D geometric model with a grid implementing an octree since humans do not carry out spatial reasoning based on a continuous map, but on a discrete one [Sow64]. The computation of the octree is achieved using the fast voxelisation method developed by [Kar99].

The goal is to introduce the equivalent of a small nervous system into the control architecture, thus linking its sensors and its effectors. Learning can modify the organization of the control architecture and that of the evolving process at the same time. The latter will be the object of our future research, as these processes are the main adaptive ones that nature has invented to ensure the survival of living beings.

An AVH is able to explore its VE and work out a mental representation of its spatial organisation in the form of a “cognitive map”. It can then use it to locate itself and reach a given goal. This ability is based on the use of a performing visual system as described before. An AVH learns a more general model of the virtual world during its interactions with the VE. This model helps it anticipate how the VE changes depending on actions that are performed.

An AVH is situated in a simulated VE with sensors for vision, audition and touch, which inform it of its external VE (active and predictive perception) or its internal state (proprioception). An AVH has effectors permitting it to act on the VE and a control architecture that coordinates its perceptions and actions. The behaviour of an AVH is adaptive as long as the control architecture allows it to maintain its essential variables in their viability zone (e.g. a corrective action was accomplished at point B, to avoid leaving the viability zone at point A). The control architecture plays the role of a motivational system when it is used to choose successive goals that the AVH is trying to reach or to arbitrate between conflicting goals.

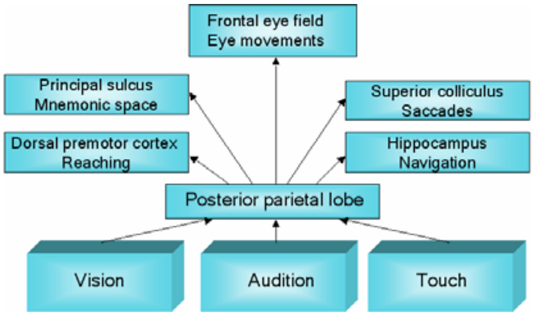


Figure 1: A schematic representation of the standard theory for multi-sensory spatial integration and sensory-motor transformations. Sensory modalities encode the location of objects in reference frames that are specific to each modality. Multi-sensory integration occurs in multiple modules with the parietal cortex.

The auditive position of an object is predicted from its visual position. This requires the transformation of a reference system whose origin is a vision coordinate (eye position) to a reference system whose origin is an audition coordinate (head position). The comparison of the results can be used to determine whether the signals from the two types of virtual sensors belong to the same object.

Our multi-sensorial approach (Figure 2) integrates the behaviour model of an AVH. The control architecture is standardised with sub-modules covering the different techniques necessary for the artificial simulation of the AVH’s behaviour.

A major problem in behavioural learning is the introduction of automatic learning techniques in multi-agent systems. It is a challenge to teach multi-agent systems how to behave, interact or get organized in order to improve their collective performances in carrying out a task.

In this context, two major obstacles are encountered:

- The choice of a learning technique relevant to the task and the level. It should allow comparison between similar problems.
- The choice of a learning protocol.

5. Realisation

5.1 Integration of Virtual Sensors

The modelling of an AVH gaining its independence with regard to its virtual representation remains an important theme in research and is very close to autonomous robotics. It helps also to understand and model human behaviour.

The AVH collects information only through the virtual sensors described earlier.

We assume that vision is the main canal of information between the AVH and its environment as indicated by the standard theory in neuroscience for multi-sensorial integration [Elf90].

The sensorial modalities update the AVH’s cognitive map to obtain a multi-sensorial mapping. For example, visual memory in the AVH's internal memory is used for a global move from point A to point B. Should obstacles be present, it would have to be replaced for a local move by direct vision of the environment.

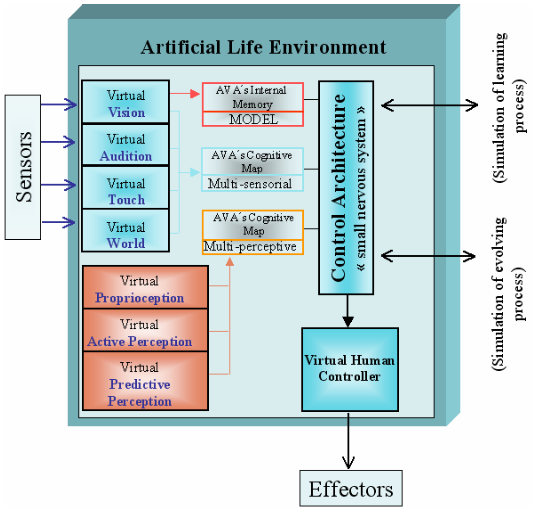


Figure 2: A schematic representation. Virtual Vision discovers the VE, constructs the different types of Perception and updates the AVH's Cognitive Map to obtain a multi-perceptive mapping. Then the Control Architecture uses both the "cognitive maps" and the "memory model" to interact with the learning, development, and control processes of the AVH (Virtual Human Controller).

In our ALifeE approach, we tried to integrate all the multi-sensorial information from the AVH's virtual sensors. In fact, an AVH in a VE may have different degrees of autonomy and different sensorial canals depending on the environment. For instance, an AVH moving in a VE represented by a well-lit room will use primarily the sensorial information of vision. However if the light is turned off, the AVH will appeal to the acoustic or tactile sensorial information in the same way a human would move around in a dark room [SKAG02].

From this observation we derive the hypotheses underlying our ALifeE framework approach. They are backed up by the latest research in neuroscience [Pou021], which describes a partial re-mapping at the behavioural level of the human including:

- *Assignment*: the prediction of the acoustic position of an object from its visual positions requires a transformation from its *eye-centred* (vision sensor) coordinates to its *head-centred* ones (auditory sensor). The comparison of these two types of results can be used to determine whether the acoustic and visual signals are directly connected to the same object.
- *Recoding*: the choice of the reference frame to integrate the sensorial signals.

Figure 3 shows the architecture of the AlifeE system [CT04].

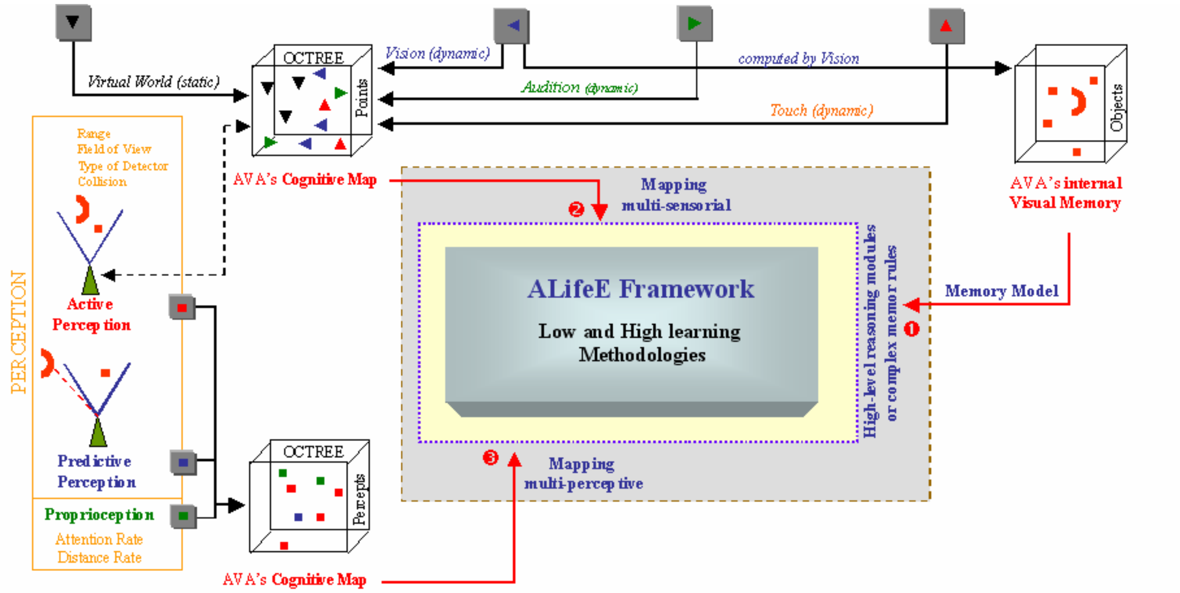


Figure 3. The architecture (*ALifeE*) used for our experimental results. The semantic information coming from *ALifeE* multi-sensorial mapping, multi-perceptive mapping and memory models is used by learning processes to establish high-level reasoning modules or complex memory rules.

References

[BBT99] Bordeux C, Boulic R, Thalmann D. An Efficient Perception Pipeline for Autonomous Agents. In Proceedings of Eurographics, 1999; 23-30.

[Ber94] Bersini H. Reinforcement Learning for Homeostatic Endogenous Variables. In Proceedings of 3rd Int. Conference On Simulated & Adaptive Behaviour, 1994;325-33, MIT Press.

[BTM96] Blumberg B, Todd P, Maes P (1996) No bad dogs: Ethological lessons for learning in hamsterdam. In: Proceedings of the 4th International Conference on the Simulation of Adaptive Behavior.

[Car02] Carollo C. Sound Propagation in 3D Environment. Ion Storm, 2002.

[CT04] T. Conde, D. Thalmann, An Artificial Life Environment for Autonomous Virtual Agents with multi-sensorial and multi-perceptive features , *Computer Animation and Virtual Worlds* , Volume 15, Issue 3-4, John Wiley, 2004

[Elf90] Elfes G. Occupancy Grid: A Stochastic Spatial Representation for Active Robot Perception. In 6th Conference on Uncertainly in AI, 1990.

[HBMT95] Huang Z, Boulic R, Magnenat Thalmann N, Thalmann D. A Multi-sensor Approach for Grasping and 3D Interaction. In Proceedings of Computer Graphics International, 1995; 235-254 Academic Press.

[HLC*97] Hudson T, Lin M, Cohen J, Gottschalk S, Manocha D. V-COLLIDE: Accelerated Collision Detection for VRML. In Proceedings of VRML, 1997; 119-125.

[Kar99] Karabassi E A. A Fast Depth-Buffer-Based Voxelization Algorithm. Journal of graphic tools, 1999; 4(4):5-10, 1999.

[KL99] Kuffner J J, Latombre J C. Fast Synthetic Vision, Memory, and Learning Models for Virtual Humans. In Proceedings of Computer Animation, IEEE, 1999; 118-127.

[Lin01] Linas R. I of the Vortex: from Neurons to self, 2001. MIT Press.

[MGD04] Monzani J.S,, Guye-Vuilleme A, de Sevin E., Behavioral Animation in : Magnenat-Thalmann & Thalmann D (2004) Handbook of Virtual Humans, John Wiley and Sons

[Mit97] Mitchell T. Machine Learning, 1997. Eds. McGraw Hill.

[NRTM95] Noser H, Renault O, Thalmann D, Magnenat Thalmann N. Navigation for Digital Actors based on Synthetic Vision, Memory and Learning. Computers and Graphics, 1995; 1:7-19.

[NT95] Noser H, Thalmann D. Synthetic Vision and Audition for Digital Actors. In Proceedings of Eurographics, 1995; 325-336.

[Pou05] Pouget A. A computational perspective on the neural basis of multi-sensory spatial representations. Nature Reviews/Neuroscience, 2002; 3:741-747.

[Rey93] Reynolds CW. An evolved, vision-based behavioral model coordinated group motion. In From Animals to Animats, 2nd International Conference on Simulation of Adaptive Behavior, 1993; 384-392. MIT Press.

[RMT90] Renault O, Magnenat-Thalmann N, Thalmann D. A Vision-based Approach to Behavioural Animation. Journal of Visualization and Computer Animation, 1990; 1:18-21.

[S02KAG] Strösslin Th, Krebser Ch, Arleo A, Gerstner W. Combining Multimodal Sensory Input for Spatial Learning. In Proceedings of ICANN, 2002; 87-92. LNCS 2415.Springer-Verlag.

[Sow64] Sowa J F. Conceptual Structures, 1964. Ed. Addison Wesley Company.

[TT94] Tu X, Terzopoulos D. Perceptual Modeling for the Behavioral Animation of Fishes. In Pacific Graphics, 1994. Ed.World Scientific.

[Wen92] Wenzel E M. Localization in Virtual Acoustics Displays. In PRESENCE, 1992; 1(1): 80-107.

[Yva02] Yvanov Y. A. State Discovery for Autonomous Learning. PhD Dissertation, MIT, 2002.

II-4: Hardware for Mixed Realities in Inhabited Worlds

Frédéric Vexo, Mario Gutierrez, Sylvain Cardin, Achille Peternier

Virtual Reality Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)

Abstract

This tutorial focuses on the hardware required for creating mixed reality environments. Applications mixing and virtual worlds need a variety of devices for data acquisition (video cameras, gaze tracking, global positioning etc.) and rendering (semi immersive and fully immersive displays, mobile devices, etc.). We present the state of the art of technologies used for acquiring information from the real world and for augmenting it by means of synthetic images, etc.. Besides the technology review we share our observations and experience with mixed reality environments applied in the contexts of virtual rehabilitation, teleoperation, surveillance and security.

Categories and Subject Descriptors (according to ACM CCS):
H.5.1 [Multimedia Information Systems]: Artificial, augmented and virtual realities
H.5.2 [User Interfaces]: Haptic I/O, Input devices and strategies, Interaction styles

1. Introduction

This tutorial presents an overview of state of the art hardware for mixed-reality applications. We classify mixed reality hardware as acquisition and rendering devices. Acquisition devices involve hardware for video acquisition, head and gaze tracking, light weight geo-localization devices, etc. Rendering devices include display technologies: miniature head mounted displays, wearable computers with 3D graphics hardware acceleration, etc. The tutorial follows a logical structure describing the technologies required to built a mixed reality application. We start with the main input: real images and explain about video acquisition hardware. Then we proceed with acquisition of user input: head, gaze, position in environment (geo-localization). Tracking user's action and position is essential for mixing the real world with virtual imagery, which is calculated as a function of user interaction. The second part of the tutorial deals with rendering hardware: technology required for displaying the mixed reality environment in an immersive or semi immersive way, through mobile or fixed displays. We cover also the novel 3D acceleration chips for mobile devices and present a general overview of 3D graphics cards for PCs. The tutorial ends with a case study illustrating the use of most of the technologies we have presented in the context of a security and surveillance system based on mixed reality.

2. Acquiring Information from the Real World

This first part of the tutorial presents an overview of the data acquisition technologies commonly involved in mixed reality applications. One of the most important devices are, of course, real-time video acquisition systems. We describe the video acquisition technologies involved in a typical mixed reality application. Video Acquisition could be complemented with some kind of motion tracking and/or servos motor. The first one is for detecting, analyzing and interpreting the user's gaze, position in the environment when the second is used to drive controlled camera. We present an overview of novel tracking technologies such as eye trackers. Mixed reality applications may require global position data (GPS). We present some examples of GPS hardware and applications, including some techniques to obtain rough estimations using existing WiFi infrastructure.

2.1. Video Acquisition

Video Cameras are the essential tools for acquiring and using real images in a mixed-reality system. In this subsection we give some hints about how to design a video acquisition system according to the needs of the application. First we will present some systems suitable for low-cost or embedded applications where volume, power and weight must be reduced to the minimum. In a second part, we will overview different systems for capturing realistically the surrounding environment. Finally we will conclude presenting other types of video camera based acquisition systems designed for singular applications such as motion capture.

Digital cameras are common hardware nowadays, they are usually classified according to the type of electronic captor they use. There are two main types of electronic captors: CCD (charge coupled device) and CMOS (complementary metal oxide semiconductor). While they are often seen as rivals, CCDs and CMOS sensors have unique strengths and weaknesses that make them appropriate to different applications. Neither is categorically superior to the other. Usually we can say that CCDs offer superior image performance (as measured in quantum efficiency and noise) and better flexibility (the sensor is set by the surrounding system), whereas CMOS offer smaller system size due to a larger scale of integration of components (more functions are embedded on the chip, e.g. color balancing and sensibility settings). A deeper comparison of both sensors can be found in [Jan02, Lit01]. In [Car02], the author presents a system for analyzing their image quality.

CMOS based cameras are suitable for light or low cost applications since the electronic sensor drivers are embedded in a single chip. Many devices like webcams [Phi] or spycams [Rai] are used for rendering low quality images. These devices are very well suited for low accuracy tracking and real time vision since the image size is appropriate to the capabilities of processing algorithms. A commercial example of CMOS camera is the module CMUcam developed at the Carnegie Mellon University [CMU]. This device is based on a CMOS pinhole camera linked to a microcontroller. The system used real time vision algorithms to track mobile elements in the environment and to send control commands to two or four servocontrollers to orientate the camera. The main disadvantage of CMOS based camera systems comes from the low quality of the recovered video stream, due to its low definition and difficulty to implement in hardware advanced treatment routines to compensate for environmental changes, e.g. variable light intensities, etc.

When the camera system must provide a realistic acquisition of real world images, the natural choice are CCD based systems. Due to the better definition and image quality they provide, these systems are mostly present in digital cameras and amateur to professional video cameras. For instance, in the surveillance application described in case study of this tutorial we are using the Panasonic NVMX500 [Pan] in order to acquire a high definition video stream to be rendered in the mixed reality environment. Most of CCD devices include optical lenses and zoom to setup the desired field of view. Professional systems can allow higher performances and full software control over the camera settings, e.g. the Uniq uc930cl [Uni]. Other systems, like motion tracking using the VICON-8™ [Vic] system, require accurate motion tracking of a set of markers. The cameras used in such systems are high definition, high refresh rate and run also under infrared illumination in order to facilitate the marker recognition.

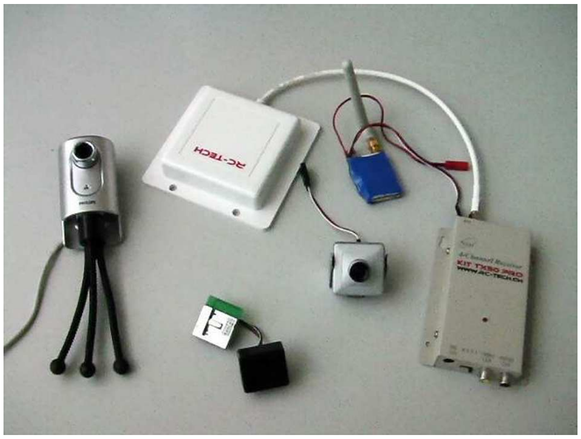


Figure 1: *Examples of cameras and video transmission systems.*

Video Transmission

Video acquired by a camera should be streamed -transmitted to a computer for treatment (incorporating synthetic images). In this sense, analogical video cameras with a composite output can be highly effective when combined with the adequate frame grabber technology. Frame grabbing consists on digitalizing an analogous video stream in order to process/visualize it in a computer application. This can be accomplished via hardware peripherals like PCI or AGP video acquisition cards. We can also cite some video grabbers, like the Cameo Grabster AV200 [Cam] that work with an USB port and allow for accessing to an analog camera like a webcam within the operating system. Despite the fact that digital video cameras are usually more efficient and directly accessible by a computer, analog systems have still their benefits, most of the wireless video transmission systems work with analogical signals due to the more compact data stream.

Concerning digital cameras, webcams (i.e. cheap cameras with no recording capabilities with USB or Firewire interfaces), are directly accessible with the help of software libraries. We can in particular cite VidCapture [Ell], an open source, efficient library for configuring the webcam retrieving its image buffer in native code. For classic digital cameras (DV, Mini-DV) with IEEE1394 (Firewire) interface, there is also a library, RapidFire [Hit], which has the same functionality as VidCapture. Usually such cameras have also a video composite output that can be used with the Cameo Grabster device described before.

The alternatives for video acquisition that we have presented consider that the video camera is directly plugged into the computer running the mixed reality application. However, in many cases we may need to place video cameras in a remote location separated from the computer performing image processing or rendering. This is frequently the case of teleoperation or outdoor mix reality applications. Our teleoperation research presented in case study section uses both software and hardware video transmission. We have used a hardware-based system for the wireless video link coming from the teleoperated vehicle to the PC controlling it. We used a $2,4Ghz$ PAL analog video emitter. We have already described how to import analog video source into a computer. The advantage of these devices is that they can transmit video over $10km$, in the line of sight. Then we have used software for sending the video stream over Internet in real-time to the operator that pilots the vehicle. Network video streaming is a complex problem that has often been addressed: we can cite RTP [SCFJ96], the Real-Time Transport Protocol, as one of the best solutions. RTP fulfills most of the requirements for network video streaming, like asynchrony solving capacity, server/client traffic regulation or multiplexing of different kinds of streams [Ben00]. Many implementations of RTP are available, but one of the easiest to deploy is the one that works in the Sun Java Media Framework (JMF) [Jav03]. The JMF-based RTP implementation is free, it contains image compression utilities useful for reducing the data bitrate (this is not in the RTP standard) and finally it is easier to program with it, even with a native programming language, than with other free libraries (we have tested LiveMedia [Liv]).

Once we are able to acquire video and process it on a computer, the next step is to acquire user interaction: motion, gestures, position in the environment. Next section describes novel technologies for Position and Gaze tracking.

2.2. Position and Gaze Tracking

Mix reality application imply the generation of synthetic images according the position of the user and his gaze orientation. This section present novel technology aimed at acquiring information about the user's gaze and his global and relative positioning.

Eye tracker

The eye is the first sense used by humans for recognition of their environment. Knowing exactly where we are looking at has been subject to research since the early 70's. Historically, the first studied application about gaze analysis was the eye typing system developed for disabled people [HTJ*04]. When the eyes are the last way to communicate with the outside world, research has been made to improve the condition of patients by giving them the ability to write text. This technique is mostly based on a virtual keyboard which is selected by the user's gaze. Focusing half a second on the desired character selects it. Many other systems have been applied, some using blinking as a Morse code, some using eye movement patterns [Iso00]. Using eye tracking is a really intuitive and accurate way to point out something. The gaze can be easily controlled and used as input. But most of the time the mechanism in charge of moving the gaze is totally unconscious. Indeed, the main applications stream of eye tracking devices is to get feedback about where the user is focusing on during a defined situation. This is a powerful tool for cognitive psychology research. A lot of applications from marketing [PHG*04], to phobic therapy [CHT04], including driving simulation [SRC*02], use gaze analysis for test and measurement purposes. Marketing applications use it to optimize placement of important information and advertisement. Therapists get some interesting data for phobia analysis by means of drawing a map of gaze attraction in the scene. On driving simulation we can measures the effect of stress and tiredness on focusing and attention.

Most of eye tracking devices are camera based only. The main difficulty for eye tracking based on camera vision is to determine the gaze vector from the image. Several systems exist in order to compute such information. Some of them, use standard webcams as detectors. The lack of accuracy inherent to the difficulty to determine the position of the eye and the pupils on low defined image generates instability of the tracking system. Even in the best conditions, when the detection of the eye and the pupils is perfect, the calculation of the gaze vector is difficult. The position of the pupils inside the eye is not always absolutely relevant. The most accurate way to compute the gaze vector is to compare images from an infrared and a standard illumination. In this case the back of the retina is visible as a bright dot on the difference image. Knowing the pupil position and the back of retina, the gaze vector can be computed more precisely assuming a standard size for the ocular globe. Some systems use fixed cameras to track the whole head of the user and compute both the gaze vector and the head position and orientation. One part of the algorithm determines the position of the head and the other the deviation of the gaze.

A system based on embedded eye tracking coupled with motion tracking for recovering the position and orientation of the head has been developed in our laboratory [CHT04], see figure 2. The eye tracking system is composed of two cameras and one infrared light source. In order to improve the accuracy, the camera is focused on the eye which covers the main part of the image. This system is used for two main purposes. The first one is to characterize the gaze attention to details present on the image. One of the applications related to psychology is to determine the gaze behavior when a patient is under stress, like a job interview for example. The second main stream of application is to use the eye tracking system as an input for pick and select a region of the image.

The eye tracking system measures the orientation angles of the gaze within 0.5° with proper calibration. This is equivalent to $3cm$ accuracy on a screen at $3m$ distance. The coupling with the motion tracker and the instability of

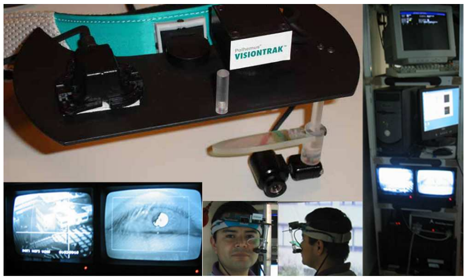


Figure 2: *The eye tracking system.*

the eye vision recognition perturbs the measurement. The calibration is crucial for recovering proper values. It has to be done before every measurement session and function of the person and the position of the head fixation. We are currently working adapting such system for a four sided CAVE describe in (see section 3.1).

We have presented an overview of motion tracking technologies working at different scales of the human body, going from main body limbs down to hands, fingers and ending with eye tracking. Next section will deal with technology aimed at tracking users or objects at a global scale: global positioning systems applied to mixed reality.

Global Positioning

The Global Positioning system (GPS) is a satellite-based navigation system made up of a network of 24 satellites placed into orbit by the U.S. Department of Defense. GPS was originally intended for military applications, but in the 1980s, the U.S. government made the system available for civilian use. Nowadays we find GPS systems integrated in cars and other public transportation means. GPS have evolved into light, transportable and affordable devices. Several mixed-reality applications for culture and entertainment make use of GPS. Global positioning is used to locate the user and her movement in the real world. The real position is then used to place the user in a virtual environment.

The Distributed Multimedia Research Group at the Lancaster University, have developed a multiplayer game based on GPS and handheld devices [MMM*03]. In [HCB*01] such system is used to simulate archeological roaming for artefact collection. The user holds a mobile device which acts as a virtual sensor to detect virtual objects in a open space. Once finished, a complement of information about the gathered artifacts is available through an interactive interface.

Today’s GPS receivers are extremely accurate, thanks to their parallel multi-channel design. Certain atmospheric factors and other sources of error can affect the accuracy of GPS receivers which is 15 meters on average. Newer GPS receivers with WAAS (Wide Area Augmentation System) capability or with Differential GPS (DGPS) can improve accuracy to less than three meters on average.

One main advantage of GPS system is the possibility to locate a point everywhere on the surface of our planet. Unfortunately, GPS-based solutions are suitable only for outdoor tracking. Furthermore, GPS accuracy may be affected by external perturbations, like atmospheric phenomena and physical occluders. As explained in next subsection, there are alternatives to GPS that allow for calculating the position of the user, when a rough estimation is enough.

WiFi access points as positioning system

WiFi nets are composed of access points allowing clients within a specific range to connect to the LAN through it. Usually, WiFi cards and HUBs allow for calculating the quality of the signal between users and access points. When several access points are in range and their coordinates are known, this information can be used to approximatively determine the user’s position. This approach is best suited for places with a good amount of access points, so that the user is always covered by more than one signal. Such technique can be useful to retrieve in which building or area a person is or, viceversa, for a user to determine his/her approximative position. For example, in a museum context, a user could download a virtual guide software on his/her PDA at the entrance and get comments and explications according to the museum’s rooms along with the visit. In comparison to GPS techniques, this method does not need specific hardware and comes for free on already WiFi equipped areas or buildings. Free software like NetStrumbler [Stu] access such information.

Applications of global positioning

At VRlab we are currently working on a project named “Flying Camera”, based on GPS technology. The main goal is to use a blimp to carry different video cameras which can be used in the context of mixed-reality applications: augmenting video stream with virtual objects/characters. Moreover, global positioning is essential for teleoperation. Figure 3 shows a 2D view of the surveyed site and the current position of the blimp. We use the STXm-900 OEM module from GPSFlight [GPS]. This device is able to simultaneously record position data to local memory for remote retrieval over an air link and transmit telemetry over a long range wireless data link in real-time. The receiver is connected to the computer via USB2.0 considered as a COM Port (RS-232). We obtain data about location, speed and pressure to provide a 2D/3D view of the terrain (see figure) being surveyed by a teleoperated blimp which carries on the GPS emitter.

Summary

We have presented some of the main technologies involved in the acquisition of data useful for mixed-reality applications: visuals from the real world (video acquisition),

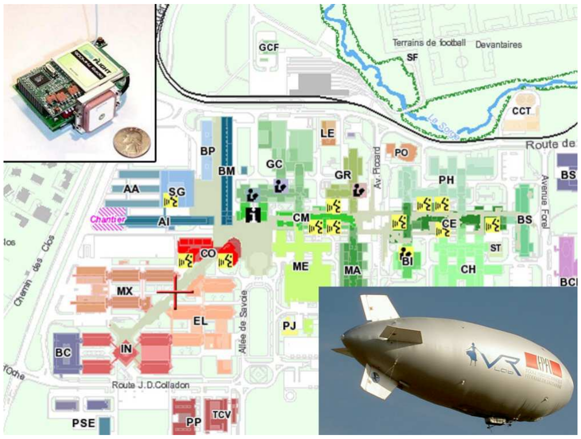


Figure 3: *Using GPS for visualizing the location of a tele-operated blimp.*

global position of users and objects (GPS), motion tracking of user's movements at the level of the eye. Next section will describe the different kinds of hardware for mixing -rendering- virtual objects over real scenes.

3. Mixing the Virtual with the Real

This section is the complement to the technologies presented in the first part of this tutorial. Data acquisition devices are used in combination with rendering technologies for mixing the reality with virtual images. The main rendering technologies we overview are concern imaging: augmenting reality with 3D graphics. We consider that image rendering can be done in different contexts that depend on the level of immersion and mobility of the user. The tutorial covers image rendering and display hardware for fully immersive, semi immersive and mobile mixed reality applications.

3.1. Image Rendering

In this section we present an overview of the current state of the art of rendering systems, focusing on immersive/semi-immersive systems capable to display stereographic images in mixed-reality contexts and Mobile devices endowed with 3D rendering capablities.We also present the way we built our own low-cost CAVE.

Standard PCs

Thanks to the entertainment industry, the 3D capabilities of home PCs have incessantly grown since the last few years. Low-cost graphic cards (like the ones produced by NVidia [NVi] or ATI [ATI]) offer today almost the same functionalities envn more that once offered by professional, specific and expensive hardware (such as SGI graphics workstations [Sil]). Complete and well documented programming libraries (OpenGL, DirectX) allow the user to interface and benefit of such 3D power. Moreover, last generation 3D graphic cards come with a customizable pipeline: that is, the user can replace the core of the hardware accelerated rendering process (like triangle filling and pixel plotting) with customized instructions coded through specific programming languages (like CG, HLSL or GLSL). This leads to a new level of freedom in the field of computer graphics applications and distribution.

Unfortunately, desktop PC displays are not well designed for immersive virtual reality applications nor for stereographic rendering, although stereovision on high frequencies desktop CRT-screens is available through wearable shutter glasses (see section 3.1) or by using a 3D desktop screen (like Stereographics' SynthaGram [Ste]). Despite of this limitation, low-cost home PC graphic cards can be used with head mounted displays or as rendering machines for wall displays or customized CAVE installations.

Mobile devices

Pocket PCs, mobile phones and mobile gaming devices are also experiencing an increasing improvement of their 2D/3D graphic capabilities: computing time expensive tasks like texture mapping and filtering, high-resolution meshes or hardware transform and lighting rendering are now available on recent handheld devices.

For the moment, just some models come with a true embedded 3D graphics chip (like PowerVR's MBX Lite [Pow], ATI's Imageon suite [ATI] or NVidia's GoForce serie [NVi]). PowerVR's MBX Lite leads this category and comes with some interesting solutions to speedup rendering times and reduces resources and battery consumption. It uses a hardware tile-based rendering system which determines the needed amount of updates on specific screen portions and allows deferred texturing (thus reducing pixel fill-rate). Hardware texture compression, 24bit true colors, transform and lighting, multi-texturing, z-buffering and performance-lossless full screen anti-aliasing are just some of the functionalities this chip features. Such functionalities were until now reserved to desktop PCs or gaming devices only. Furthermore, these chips are usually coupled with an embedded MPEG decoder and a JPEG codec as well, for fast image and video streaming. Moreover, mobile devices are often (or can easily be) equipped with image acquisition hardware (photo or video cameras), resulting in an ideal and non expensive platform for mobile mixed reality (this is explored in [GKRS01] and [WPLS05]).

PDAs

Pocket PCs (like Dell Axim x50v [Del]) are increasingly being equipped with 2D/3D acceleration chips. The computation power of such devices and their memory and connectivity capabilities as well (USB adapters, GPS interfaces) allow new exploitations of PDAs in the computer graphics

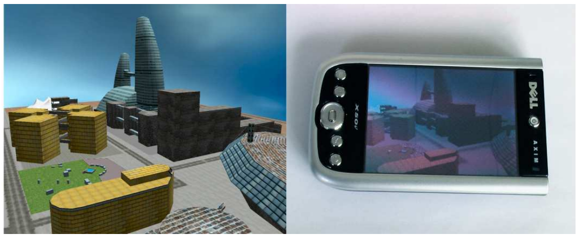


Figure 4: *Dell Axim x50v displaying a highly detailed 3D scene.*

context. Furthermore, high-end PDAs can be connected to external displays through a VGA port: head-mounted and see-through head-mounted displays can be easily interfaced and used in new contexts (like outdoor and urban environments) without the limitations caused by wires and heavy notebooks.

We are currently developing at VRlab a 3D graphics engine for a Dell Axim x50v Pocket PC [Del]. This device is equipped with a Intel's 2700G multimedia accelerator [Int] which provides 3D rendering capabilities comparable to home PCs of few years ago. The Intel's 2700G includes PowerVR's MBX Lite core for 3D graphic operations. We used OpenGL ES (Open Graphic Library for Embedded System [Ope]) as software interface to the 3D accelerated functionalities. Even if the project is still in progress, we can already display complex textured models at interactive frame rates. The scene used of figure 4 is composed of 7000 triangles and 20 textures of 128x128 texels. With texture smoothing filter activated, Gouraud shading and one omnidirectional dynamic light source we obtain an average of 5 images per second at 320x240 pixels of screen resolution. We just started to implement more hardware specific optimizations which should significantly improve the frame rate.

Mobile phones

Even if they're more limited in comparison to PDAs, mobile phones also offer interesting computation power and graphic functionalities (like Nokia's NGage [Nok]). Most of the principles discussed in the PDA section can be applied to mobile phones as well.

Portable gaming devices

Portable gaming devices are now showing stunning 3D features. Nintendo's Dual Screen [Nin] portable console comes with two screens (one with touch detection, like a PDA) and offer high-quality 3D rendering. PSP [Son] shows also impressive 3D rendering features. Unfortunately, the hardware architecture and the software packages of such devices, due to proprietary constraints, are less available and versatile in comparison to a PDA.

PDAs, mobile phones and mobile gaming devices have recently acquired enough computation power and versatility to be considered as interesting platforms and tools for VR/AR. They offer 3D realtime rendering and programming capabilities with affordable costs with a wearable format. When coupled with semi-immersive or immersive displays (external screens, HMDs, see-through HMDs) or with image acquisition devices (photo or video cameras) they can be used in mixed reality contexts not only constraint to small rooms or CAVEs. Furthermore, GPS systems can be added to allow a virtually worldwide extend to such contexts.

Head-Mounted Displays and See-through HMDs

Head-mounted displays (HMDs) are one of the most common ways to provide fully immersive experiences. Since binocular vision considerably enhances visual depth perception, HMDs usually come with separate screens for right and left eye thus allowing stereoscopic vision (see figure 5). Special optics in front of the screens guarantee a wide field of view. A head tracking system helps to locate the user's head position and orientation in real time.

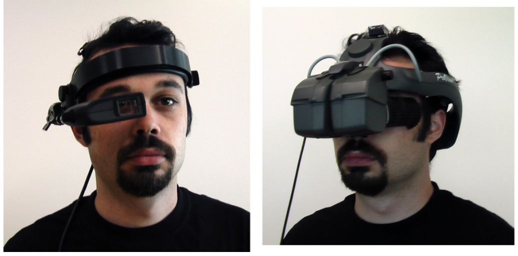


Figure 5: *Head Mounted Displays.*

See-through head-mounted displays are a special case of standard HMDs. They use semi-transparent/transparent screens on which images are displayed (like the ones used in combat aircrafts to show current information directly on the cockpit's glass). See-through HMDs, can mix synthetic images with the real world behind the transparent displays and are therefore extensively used in mixed-reality applications. Some less encumbering see-through HMDs use laser projection to render images directly onto the retina [JW95].

Non-transparent displaying systems tend to confine the user in the virtual world avoiding contact with the real world. With such devices, mixed-reality can be achieved only by placing a camera on the user's head and by using the video stream as background image on both screens: an even unpractical solution if stereovision (and thus two separate cameras) is required.

We experienced that displaying a video stream coming from the real world within a virtual environment helps users to feel more comfortable. During a demo scene (see figure 6), we put a camera near our virtual reality installation. When a visitor wanted to try our system (based on a non see-through HMD), the same person who helped him wearing the VR suite was still present in the virtual world on a 3D television, keeping a contact with the user and thus partially bypassing the lack of interaction and reducing the absence of cooperation an HMD immersion usually involves.

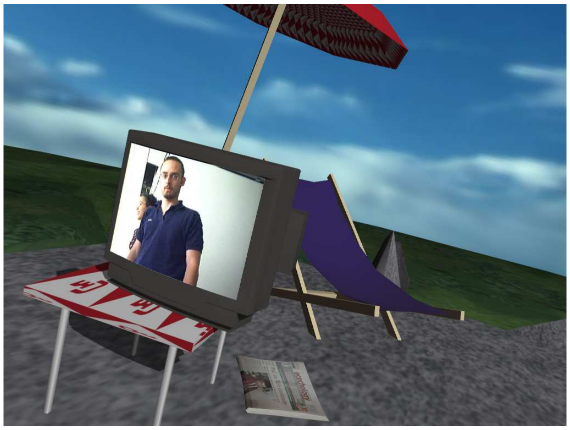


Figure 6: *Virtual Environment mixed with live video streaming.*

Shutter Glasses

The shutters are wearable glasses synchronized with a display device which alternately open and close at 120*Hz* in conjunction with the alternating display of the left and right eye view on the display, presenting each eye with an effective 60*Hz* refresh. They allow for showing a unique and different image to each eye. As a result, the brain integrates these two views into a stereo picture. An infrared emitter is charged of sending synchronization signals to one or more shutters. Usually, they are used with a wall display or a CAVE installation, but low budget solutions exist for home PCs with a 120*Hz* CRT monitor too.

Wall displays

A wall display may be considered as a big cinema display. Stereoscopic imaging can be achieved as well by using shutter glasses worn by users and a high refresh rate CRT beamer ($100-120Hz$) or two LCD/standard projectors with shutters. A wall display permits semi-immersive virtual reality if users are close enough to the screen. Wall displays offer the opportunity to have an interesting immersive context while keeping the user body in the virtual space rather than his/her eyes only. Moreover, wall displays allow the simultaneous use of an eye tracking system which is impossible with a normal HMD.

CAVE

The CAVE Automatic Virtual Environment (or simply CAVE) was originally conceived in 1991 by Thomas DeFanti and Dan Sandin and implemented by Carolina Cruz-Neira at the University of Illinois at Chicago [CNLP*93]. The idea behind this project was to create a VR system without the common limitations of previous VR solutions, like poor image resolution, inability to share the experience directly with other users and the isolation from the real word. Physically, a CAVE is composed by a cube of display-screens that surround the viewer. Those screens have stereoscopic images projected on and give the illusion to the users to be completely immersed in the virtual world. A head and hand tracking system is used to produce the correct stereo perspective and to identify the position and orientation of a tree-dimensional input device within the scene: this lets the user see his or her entire environment from the correct perspective, thus creating a compelling illusion of reality. Real and virtual objects are blended in the same space and the user can see his/her body interacting with the environment. Audio interaction can also be used in a CAVE simulation. A CAVE is usually composed of three to six projection screens. Some authors refer under the term “CAVE” even dual or single screen solutions. Dual screen solutions are sometimes referred as “V-CAVE”, because the two screens form a letter “V”. Single screen solutions are in fact an evolution of wall display, sometimes projecting images on slightly curved surface to improve the immersive design of the system. The quality and the price of a CAVE installation largely depends on the needs and budget of the users. High-end professional implementations exist, like the ones offered by Fakespace, Pyramid Systems or SGI. Such solutions use high frequency CRT projection systems which can display images at 120*Hz* on each wall (60*Hz* with shutter glasses), typically driven by a powerful multiprocessor computer which can handle multiple monitors or video displays at the same time. Finally, the need for rendering different images for each wall requires more computational power at the graphics level in order to obtain real-time interaction with the scene. On the other hand, V-CAVEs are, a cheaper solution for experiencing with immersive virtual spaces. In fact, a V-CAVE can be built with just two beamers directly projecting the images on a bright room corner. Jacobson et. al. suggest a cost of around $2500 per screen [JKES05].

Building a low-cost CAVE

Unfortunately, the price of such kind of hardware is extremely high and rarely affordable: the final cost of a commercial CAVE may raise up to $1 million. At VRlab we have built a custom implementation of a CAVE, using widely available hardware and materials in order to drastically reduce costs.

We have used a three-wall + floor architecture for our CAVE: the front screen is 2.40 meters large and 2.20 meters high, side panels are 1.65 meters long. The interior of our installation can host comfortably up to three persons. We used four LCD projectors with up to 75 hz of refresh rate. Such hardware offers a good overall luminosity. In the future, we will add four more projectors to implement stereographic rendering, using one beamer for each eye and by synchronizing the images through shutter glasses for users and shutter panels mounted on LCD projectors. Synchronization signals will be broadcasted by an external shared synchronization box. Those beamers are separately driven by standard PCs powered with last generation graphic cards for home computers. The dual screen support of those graphic adapters (widely available) is used to connect two projectors to each PC. A fifth PC acts as server, managing the synchronization between the four rendering machines and handling the overall state of the simulation. Gigabit Ethernet adapters are used to assure a fast and latency-proof binding between the CAVE's computers. The final cost of our implementation will be between 5–10% of a standard CAVE price.

Wall displays and CAVEs are the current trend of immersive VR/AR experiences. They are physically non-invasive for the user and they allow for implementing shared virtual environments, high resolution stereographic images and direct interaction between the user's body and the virtual world. In comparison, the advantage of using a HMD is its mobility. When a user wearing a HMD and a wearable computer is moving around in the real world, the pose of the user's head is tracked in real time by the camera of the HMD. The object recognition and the graphic image rendering are accomplished by the wearable computer. That technology is more useful in outdoor applications where a graphics workstation is non-reachable.

Our low cost CAVE offers a good compromise between the high quality and precision of a professional installation and the affordable requirements of a customized one. When compared to professional CAVEs, we experienced some problems with our installation. The front, side and floor displays have different sizes and their beamers are placed at different distances, due to space constraints. This leads to different pixel sizes and luminosity when images are projected on the walls, and requires some adjustments at rendering level. Moreover, the beamers are not conceived to perfectly fit the screen sizes, so some rays are projected outside the wall surfaces. This problem can be easily solved by applying some masks either via software (by simply adding an alpha channelled texture with black pixels on the corners and transparent in the center) or via hardware (by superimposing an occlusion surface on the beamer's lenses).

4. Conclusion

At this point we finish our overview of hardware and technology for acquiring information from the real world and mixing -rendering- it with virtual stimuli addressing the main human senses. In the next part of this tutorial we will describe a full mixed-reality case-study involving several data acquisition and rendering technologies: : A mixed-reality system for surveillance and security.

References

[ATI] ATI Corporation: http://www.ati.com/. 5

[Ben00] Benslimane A.: Real-time multimedia services over Internet. In *Proceedings of the 1st European Conference on Universal Multiservice Networks, ECUMN'00* (2000), pp. 253–261. 3

[Cam] Cameo Grabster AV200: Terratec, video-editing products http://videoen.terratec.net. 2

[Car02] Carlson B. S.: Comparison of modern CCD and CMOS image sensor technologies and systems for low resolution imaging. In *Proceedings of IEEE Sensors* (2002), vol. 1, pp. 171–176. 2

[CHT04] Ciger J., Herbelin B., Thalmann D.: Evaluation of gaze tracking technology for social interaction in virtual environments. In *Proceedings of Workshop on Modelling and Motion Capture Techniques for Virtual Environments (CAPTECH 2004)* (2004), pp. 67–72. 3

[CMU] CMUcam Vision Sensors: http://www-2.cs.cmu.edu/ cmucam/. 2

[CNLP*93] Cruz-Neira C., Leigh J., Papka M., Barnes C., Cohen S. M., Das S., Engelmann R., Hudson R., Roy T., Siegel L., Vasilakis C., DeFanti A., Sandin D. J.: Scientists in wonderland: A report on visualization applications in the cave virtual reality environment. In *Proceedings of IEEE Symposium on Research Frontiers* (1993), pp. 59–66. 7

[Del] Dell: http://www.dell.com/. 5, 6

[Ell] Ellison M.: VidCapture, Simplified Video Capture for Web Cameras, Codevis Project, http://www.codevis.com/vidcapture/. 2

[GKRS01] Geiger C., Kleinnjohann B., Reimann C., Stichling D.: Mobile AR4ALL. In *Proceedings of the Second IEEE and ACM International Symposium on Augmented Reality (ISAR)* (2001), pp. 181–182. 5

[GPS] GPSFlight, Inc.: http://www.gpsflight.com. 4

[HCB*01] Hall T., Ciolfi L., Bannon L., Fraser M., Benford S., Bowers J., C.Greenhalgh, Hellström S.-O., Izadi S., Schnädelbach H., Flintham M.: The visitor as virtual archaeologist: Explorations in mixed reality technology to enhance educational and social interaction in the museum.

In *In Proceedings of VAST 2001, Conference on Virtual Reality, Archaeology and Cultural Heritage* (November 2001), ACM Press, pp. 91–96. 4

[Hit] HITCHCOCK-MANTHEY, LLC: RapidFire Video Library, the Live Video Solution, http://www.hitmaninc.com/rfLibrary2.html. 2

[HTJ*04] HANSEN J. P., TØRNING K., JOHANSEN A. S., ITOH K., AOKI H.: Gaze typing compared with input by head and hand. In *Proceedings of the symposium on Eye tracking research and applications* (2004), pp. 131 – 138. 3

[Int] INTEL CORPORATION: http://www.intel.com/. 6

[Iso00] ISOKOSKI P.: Text Input Methods for Eye Trackers Using Off-Screen Targets. In *Proceedings of the symposium on Eye tracking research and applications* (2000), pp. 15 – 21. 3

[Jan02] JANESICK J.: Dueling Detectors: CCD or CMOS? *OE Magazine* (February 2002), 30–33. 2

[Jav03] JAVA MEDIA FRAMEWORK: Sun Microsystems, http://java.sun.com/products/java-media/jmf/, 2003. 3

[JKES05] JACOBSON J., KELLEY M., ELLIS S., SEETHALER L.: Immersive displays for education using caveut. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (2005). 7

[JW95] JOHNSTON R., WILLEY S.: Development of a commercial virtual retinal display. In *Stephens, W. and Haworth, L. A. (Eds.) Proceedings of Helmet and Head/mounted displays and symbology design* (1995), pp. 2–13. 6

[Lit01] LITWILLER D.: CCD vs. CMOS: Facts and Fiction. *Photonics Spectra* (January 2001). 2

[Liv] LIVEMEDIA: Live.com, streaming media http://www.live.com/livemedia/. 3

[MMM*03] MITCHELL K., MCCAFFERY D., METAXAS G., FINNEY J., SCHMID S., SCOTT A.: Six in the city: introducing real tournament - a mobile ipv6 based context-aware multiplayer game. In *NETGAMES '03: Proceedings of the 2nd workshop on Network and system support for games* (New York, NY, USA, 2003), ACM Press, pp. 91–100. 4

[Nin] NINTENDO: http://www.nintendo.com/. 6

[Nok] NOKIA: http://www.nokia.com/. 6

[NVi] NVIDIA CORPORATION: http://www.nvidia.com/. 5

[Ope] OPENGL ES, KHRONOS GROUP: http://www.khronos.org/. 6

[Pan] PANASONIC: http://www.panasonic.com/. 2

[PHG*04] PAN B., HEMBROOKE H. A., GAY G. K., GRANKA L. A., FEUSNER M. K., NEWMAN J. K.: The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the symposium on Eye tracking research and applications* (2004), pp. 147 – 154. 3

[Phi] PHILIPS: http://www.consumer.philips.com. 2

[Pow] POWERVR: http://www.powervr.com/. 5

[Rai] RAIDEN TECH INC.: http://www.raidentech.com/miwispyca.html. 2

[SCFJ96] SCHULZRINNE H., CASNER S., FREDERICK R., JACOBSON V.: RTP: A Transport Protocol for Real-Time Applications. In *Request for Comment RFC-1889 of the Internet Engineering Task Force, IETF* (January 1996). 3

[Sil] SILICON GRAPHICS CORPORATION: http://www.sgi.com/. 5

[Son] SONY: http://www.sony.com/. 6

[SRC*02] SODHI M., REIMER B., COHEN J. L., VASTENBURG E., KAARS R., KIRSCHENBAUM S.: On-Road Driver Eye Movement Tracking Using Head-Mounted Devices. In *Proceedings of the symposium on Eye tracking research and applications* (2002), pp. 61 – 68. 3

[Ste] STEREOGRAPHICS CORPORATION: http://www.stereographics.com/. 5

[Stu] STUMBLER DOT NET: http://www.stumbler.net/. 4

[Uni] UNIQ VISION INC.: http://www.uniqvision.com/. 2

[Vic] VICON MOTION SYSTEMS: http://www.vicon.com/. 2

[WPLS05] WAGNER D., PINTARIC T., LEDERMANN F., SCHMALSTIEG D.: Towards massively multi-user augmented reality on handheld devices. In *Proceedings of the Third International Conference on Pervasive Computing (Pervarsive)* (2005). 5

III: Simulating Life in a Mixed Realities Pompeii World

Nadia Magnenat-Thalmann, George Papagiannakis

MIRALab, University of Geneva, Geneva, Switzerland

Abstract

we describe a complete methodology for real-time integrated mixed reality systems that feature realistic complete simulations of animated virtual human actors (clothes, body, skin, face) who augment real environments and re-enact staged storytelling dramas. Although initially targeted at Cultural Heritage Sites, the paradigm is by no means limited to such subjects. The abandonment of traditional concepts of static cultural artifacts or rigid geometrical and textual augmentations with interactive, augmented historical character-based event representations in a mobile and wearable setup, is the main contribution of the described work as well as the proposed extensions to AR Enabling technologies: a VR/AR character simulation kernel framework with character to object interaction coupled with a markerless camera tracker specialized for non-invasive geometrical registration on heritage sites. We demonstrate a real-time case study on the actual site of ancient Pompeii.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]:

1. Introduction

Mixed Realities [MK94] and their concept of cyber-real space interplay invoke such interactive digital narratives that promote new patterns of understanding. However, the "narrative” part, which refers to a set of events happening during a certain period of time and providing aesthetic, dramaturgical and emotional elements, objects and attitudes ([NM00], [TYK01]) is still an early topic of research. Mixing such aesthetic ambiences with virtual character augmentations [CMM*03] and adding dramatic tension has developed very recently these narrative patterns into an exciting new edutainment medium [LHM03]. Since recently, AR Systems had various difficulties to manage such a time-travel in a fully interactive manner, due to hardware & software complexities in AR ‘Enabling Technologies’ [ABB*01]. Generally the setup of such systems was only operational in specific places (indoors-outdoors) or with specific objects which were used for training purposes rendering them not easily applicable in different sites. Furthermore, almost none of these systems feature full real-time virtual human simulation. With our approach, based on an efficient real-time tracking system, which require only a small pre-recorded sequence as a database, we can setup the AR experience with animated virtual humans anywhere, quickly. With the interplay of a modern real-time framework for integrated interactive virtual character simulation, we can enhance the experience with full virtual character simulations. Even if the environmental conditions are drastically altered, thus causing problems for the real-time camera tracker, we can re-train the camera tracker to allow it to continue its operation.

The proposed set of algorithms and methodologies aim to extend the “AR Enabling Technologies” in order to further support real-time, mobile, dramaturgical and behavioured Mixed Reality simulations, as opposed to static annotations or rigid geometrical objects. Figure 1 depicts fully simulated virtual humans (skin, clothes, face, body) augmenting a cultural heritage site.

1.1 Overview

As a preprocessing stage, our real-time markerless camera tracker system is being trained on the scene that is aimed to act as the mixed Reality stage for the Virtual actors. During real-time mobile operation and having already prepared the VR content for the virtual play, our system allows the user to be immersed fully in the augmented scene and for the first time witness storytelling experiences enacted by realistic virtual humans in mixed reality worlds. The minimal technical skills and hardware configuration required to use the system, which is based on portable wearable devices, allow for easy setup in various indoors and outdoors feature rich locations. Thus in Section 2 of this work we review the previous work performed in the main areas of "AR Enabling technologies", such as camera tracking and illumination as well as the extensions that we propose: complete VR character simulation framework with character to object interactions as well as a new illumination model for MR. In Section 3 we present such a framework which is mandatory in order to handle the exponential complexity of virtual character drama, that traditional rendering-centric AR systems cannot anymore handle. Finally in section 5 we present our two case studies in a controlled environment as well as on the site of ancient Pompeii and we epitomize with the discussion and conclusions sections 6 and 7 respectively.

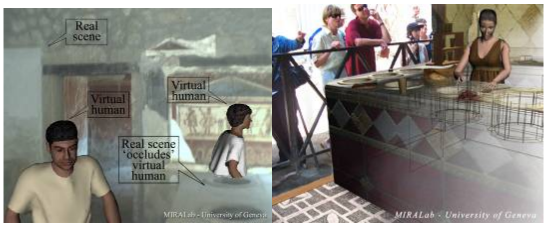

Real scene
Virtual human
Virtual human
Real scene 'occludes' virtual human
MIRALab - University of Geneva
MIRALab - University of Geneva

Figure 1. Example of mixed reality animated characters acting a storytelling drama on the site of ancient Pompeii (view from the mobile AR-life system i-glasses)

2. Related Work

On AR integrated platforms, a number of projects are currently exploring a variety of applications in different domains such as medical [ART*05], cultural heritage [SDS*01], [PSO*05], training and maintenance [WT00] and games [TDS*00]. Special focus has recently been applied to system design and architecture in order to provide the various AR enabling technologies a framework [GHJ*94] for proper collaboration and interplay. Azuma [ABB*01] describes an extensive bibliography on current state-of-the-art AR systems & frameworks. However, few of these systems take the modern approach that a realistic mixed reality application, rich in AR virtual character experiences, should be based on a complete VR Framework (featuring game-engine like components) with the addition of the "AR enabling Technologies" like a) Real-time Camera Tracking b) AR Displays and interfaces c) Registration and Calibration. Virtual characters were also used in the MR-Project [TYK01] where a complete VR/AR framework for Mixed Reality applications had been created. Apart from the custom tracking/rendering modules a specialized video and see-through HMD has been devised. However, none of the aforementioned AR systems can achieve to date, realistic, complete virtual human simulation in AR featuring skeletal animation, skin deformation, facial-speech and clothes simulation. For realizing the dynamic notions of character based Augmented Heritage, the above features are a prerequisite.

Camera tracking methods can be broadly divided into outside-in and inside-out approaches, depending on whether the sensing device is located on the tracked object, or multiple sensing devices surround the tracked object. Technologies used to perform the tracking include mechanical [Rob05], magnetic [ASC05], optical [VIC05], inertial [INT05], ultrasound or hybrids [FHP98] of these. Due to the sensitive nature of the environment our tracking system had to function in, namely the ancient ruins of Pompeii, we opted to develop an inside-out optical system without the need for fiducial markers. Visual through-the-lens tracking is a widely researched topic and several papers have been published investigating different methods of achieving this goal. A software library has also been released called ARToolKit [ART*05] which is widely used in the Augmented Reality community. This relies on large fiducials placed throughout the scene which are identified by the system and used to perform pose estimation. [DK01] has published several papers based on Simultaneous Localization and Mapping (SLAM), a very promising real-time probabilistic visual method of tracking. Another popular approach is that reported by [VLF04], which is based on a prior scene model with real-time line and texture matching. The visual fiducial method commercialized by Radamec [TJN*97] has proven to be very accurate but has the disadvantage of requiring severe scene modifications and extensive setup. Our real-time markerless tracking method has been already described in [PSO*05].

3. MR Framework components for character simulation

3.1 AR-Life system design

Our AR-Life system is based on the VHD++ [PPM*03], component-based framework engine which allows quick prototyping of VR-AR applications featuring integrated real-time virtual character simulation technologies, depicted in Figure 2. The framework has borrowed extensive know-how from previous platforms such as presented by [SBT*99]. The key innovation is focused in the area of component-based framework that allows the plug-and-play of different heterogeneous human simulation technologies such as: Real-time character rendering in AR (supporting real-virtual occlusions), real-time camera tracking, facial simulation and speech, body animation with skinning, 3D sound, cloth simulation and behavioral scripting of actions.

The integrated to the AR framework tracking component is based on a two-stage approach. Firstly the system uses a recorded sequence of the operating environment in order to train the recognition module. The recognition module contains a database with invariant feature descriptors for the entire scene. The runtime module then recognizes features in scenes by comparing them to entries in its scene database. By combining many of these recognized features it calculates the location of the camera and thus the user position and orientation in the operating environment. The main design principle was to maximize the flexibility while keeping excellent real-time performance. The different components may be grouped into the two following main categories:

1. System kernel components responsible for the interactive real-time simulation initialization and execution.
2. Interaction components driving external VR devices and providing various GUIs allowing for interactive scenario authoring, triggering and control.

Finally the content to be created and used by the system was specified, which may be classified into the two following main categories: a) Static and b) Dynamic content building blocks such as models of the 3D scenes, virtual humans, objects, animations, behaviors, speech, sounds, python scripts, etc.

3.2. MR Framework operation for character simulation

The software architecture is composed of multiple software components called services, as their responsibilities are clearly defined. They have to take care of rendering of 3D simulation scenes and sound, processing inputs from the external VR devices, animation of the 3D models and in particular complex animation of virtual human models including skeleton animation and respective skin and cloth deformation. They are also responsible for maintenance of the consistent simulation and interactive scenario state that can be modified with python scripts at run-time. To keep good performance, the system utilized four threads. One thread is used to manage the updates of all the services that we need to compute, such as human animation, cloth simulation or voice (sound) management. A second thread is used for the 3D renderer, who obtains information from the current scenegraph about the objects that must be drawn as well as the image received from the camera. It will change the model view matrix accordingly to the value provide by the tracker. The third thread has the responsibility of capturing and tracking images. The last thread is the python interpreter, which allows us to create scripts for manipulating our application at the system level, such as generating behaviors for the human actions (key-frame animation, voice, navigation).

The AR system presented in **Figure 2** features immersive real-time interactive simulation supplied with proper information in course of the simulation. That is why content components are much diversified and thus their development is extremely laborious process involving long and complex data processing pipelines, multiple recording technologies, various design tools and custom made software. The various 3D models to be included in the virtual environments like virtual human or auxiliary objects have to be created manually by 3D designers. The creation of virtual humans require to record motion captured data for realistic skeletal animations as well as a database of real gestures for facial animations. Sound environments, including voice acting, need to be recorded in advance based on the story-board. For each particular scenario, dedicated system configuration data specifying system operational parameters, parameters of the physical environment and parameters of the VR devices used have to be defined as well as scripts defining atomic behaviors of simulation elements, in particular virtual humans. These scripts can modify any data in use by the current simulation in real-time. This allows us to continue running the simulation whilst some modifications are performed.

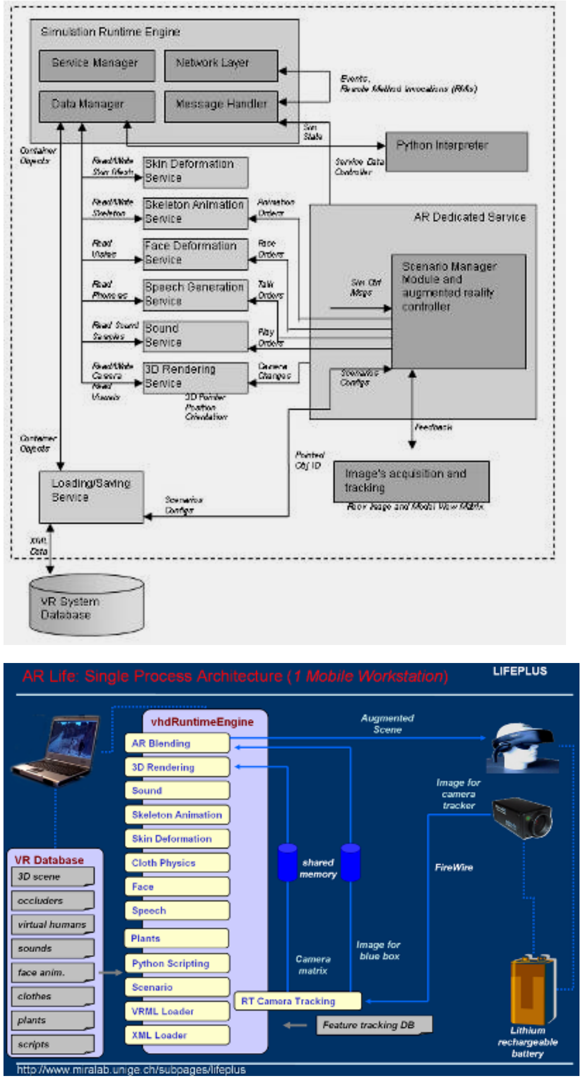


Figure 2. VHD++ AR Framework Overview

3.3. MR Registration and Staging

Employing a markerless camera tracking solution for registering the CG camera according to the real one, is an added value advantage since it eliminates the use of external tracking devices or avoids polluting the real scene with the use of known fiducial markers. However, the issue that arises is how to geometrically calibrate the camera and define the scene fiducial origin in world coordinates. Especially as our MR scenes have animated virtual characters, initial character staging, scaling and orientation is a crucial factor in order to determine correct initial, life-sized, believable geometrical registration. In the pipeline described in section 4, boujou allows for an initial scene origin to be defined offline on a tracked scene feature. This feature though is not sufficient as for a number of characters and a storytelling scenario, designers would like to interactively direct, stage and adjust the action in real-time, according to their dramaturgical interest. Therefore we propose a simple algorithm for determining the storytelling scene origin and orientation, harnessing the features of the underlying OpenGL scenegraph renderer camera metaphor and world node coordinates as depicted in Figure 3. We allow for interactive manipulation of the scene camera as well as separate scene global repositioning and scaling according to the standard OpenGL formulas for ModelView and Projection matrices: The ModelView and Projection matrices are used to set up the virtual camera metaphor and are provided in real-time for each tracked frame, by the underlying camera tracker [PSO*05]. Thus for interactive authoring-staging in the MR scene, two more controls are supplied: a) A single vector is mapped via keyboard contols on the translation part of the camera so that the camera can be furthermore tweaked within the tracked frame and b) A single 4x4 translation matrix is mapped as a virtual trackball metaphor, so that a designer can interactively stage the "Main Scene" scenegraph node that contains all the virtual augmentation. Since the basic renderer of the VHD++ framework is based on OpenScenegraph, all individual elements can have their transformation matrix modified. However, for the MR real-time authoring stage, it is important that the whole staged experience can be initially positioned according to the real scene so that the camera tracking module can subsequently register it accordingly with the real scene. For the final geometrical registration, the following algorithm was employed, according to Figure 2 and Figure 3 to modify the scenegraph virtual parts as shown Figure 3:

1. Retrieve the camera image
2. Run the feature tracker on this image
3. Extract ModelView and Projection camera matrices
4. modify the combined camera matrix according to user authoring scaling/position controls (mapped virtual trackball mouse metaphor operation)
5. apply the combined/adjusted camera matrix to scenegraph renderer
6. move Occluder geometries as root nodes in the scenegraph with GL-DEPTH test set to OFF
7. set the background image acquired from the camera in a 2D projected screen textured quad (thus since background image applied as a texture it is hardware accelerated and independent of window resolution, as hardware extension for non power of 2 textures was employed)
8. modify the Main Scene root node according to user preference for further global positioning and scaling
9. execute draw threads on the whole scenegraph

Thus with the above simple and fast algorithm we were able to both stage our fiducial MR characters and scene Occluders without any performance drop in frame rate.

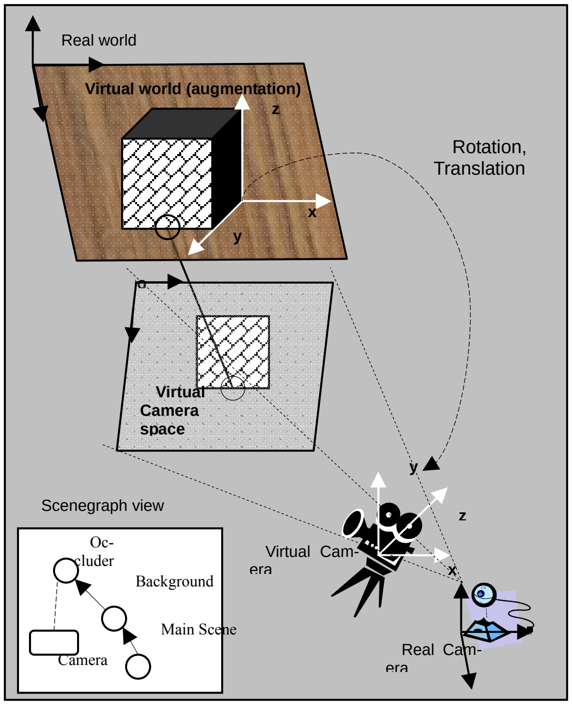


Figure 3 Camera Coordinate Systems in MR

4. Case studies and results

To meet the hardware requirements of this aim, a single Alienware P4 3.20 GHz Area-51 Mobile Workstation was used, with a GeforceFX5600Go NVIDIA graphics card, an IEEE1394 Unibrain Camera [UNI05] for fast image acquisition in a video-see-through i-glasses SVGAPro [HMD05] monoscopic HMD setup, for advanced immersive simulation. Our previous efforts were based on a client-server distributed model, based on 2 mobile workstations as at the beginning of our project a single laptop could not suffice in simulating in real-time such a complex MR application. To achieve the requirement of 'true mobility', a single mobile workstation is used currently in our final demonstrations. That was rendered possible only after the recent advances on the hardware side (processor, GPU) of mobile workstations and on our further software/algorithmic improvements (sections 3, 4, 6), in the streaming image capturing and introduction of hyper-threading and GPU calculations of the MR renderer. In all our case studies we employed 5 fully simulated virtual humans (body animation, skin deformation, speech, cloth simulation) [PSO*05], 20 smart interactive objects for them, 1 python script, 1 occluder geometry and in the case of the 'lab maquette' the 3D geometry of part of the thermopolium. The case studies statistics utilizing the above hardware configuration boil down to 20fps for the camera tracker and 17fps for the main MR simulation for the 'lab maquette' trial and 13fps and 12 fps respectively for the Pompeii trial.

4.1 Controlled environment setup trial

In order to validate that our integrated AR framework for virtual character simulation operates in different environments, we have tested the system directly in the ruins of Pompeii. However, in order to further continue AR tests in a controlled lab environment, a real paper 'maquette' was constructed in order to resemble the actual Pompeii site that we visited for our first on site tests. This allowed us for extra fine tuning and improvement of our simulation and framework, without having to visit numerous times the actual site. Figure 4 (right) depicts an example of augmenting the real 'maquette'.

4.2 Pompeii and the thermopolium of Vetutius Placidus trial

With the help of the Superintendence of Pompeii [ASP05], who provided us with all necessary archaeological and historical information, we have selected the 'thermopolium' (tavern) of Vetutius Placidus and we contacted our experiments there. The results are depicted in the following Figure 4, Figure 5 where the technologies employed for simulating and authoring our virtual humans where already described in [PSO*05].

5. Discussion and future work

From the end-user point of view, the benefits of the described approach are significant. He/she can view a superposition of the real world with a virtual scene representing fully simulated living characters in real-time. As we are able to occlude with real objects the virtual augmentations, we can therefore enforce the sensation of presence by generating believable behaviors and interaction between the real and virtual objects. However, someone could argue that no 'parthenogenesis'in the Mixed reality algorithmic field is exhibited by this study; our premise is that, as shown in Section 2, fully simulated (animated and deformed in real-time) virtual human actors have not yet appeared so far in real-time Augmented Reality providing a complete, mobile-wearable, integrated methodology that can be re-applied on multiple AR contexts with a markerless tracked camera in real-time. Our key contribution has been into providing an exciting integrated methodology-application for AR with incremental research and extensions in the 'AR Enabling technologies' [ABB*01]. Finally, there are a number of issues that need to be further improved as future work. For example the illumination registration issue is still not yet addressed which will help for rendering more realistic the MR experience and we are currently experimenting with HDRI. The real-time camera tracking performance can be further improved as well as the training process to be further shortened. Currently if the lighting conditions on the real scene are altered, the system has to be retrained and a new database to be generated. In the virtual human simulation domain, an important aspect that we will be addressing in the future is interactivity between the real users and virtual actors.

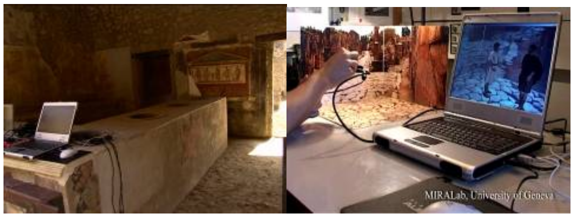


Figure 4 (left). The Real Pompeian 'thermopolium' that was augmented with virtual animated virtual characters. In this figure the scene is set-up for camera tracking preprocessing; consequently the laptop is put in the backpack for the run phase. (right) Lab 'maquette' controlled AR tests. For optimum flexibility and tests the camera is detached from the HMD and moved freely within the 'tracked area' augmenting it with virtual characters (laptop monitor)

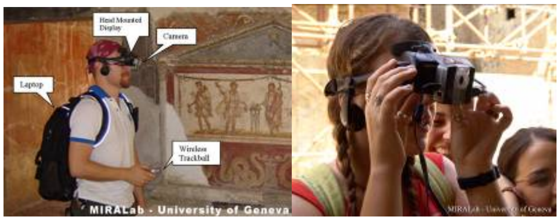

Head Mounted Display
Camera
Laptop
Wireless Trackball
MIRALab - University of Geneva
MIRALab - University of Geneva

Figure 5. The Mobile AR-Life simulator system on the actual site of Pompeii

6. Conclusions

Nowadays, when laymen visit some cultural heritage site, generally, they cannot fully grasp the ancient vibrant life that used to be integrated in the present ancient ruins. This is particularly true with ruins such as the ancient city of Pompeii, where we would like to observe and understand the behaviors and social patterns of living people from ancient Roman times, superimposed in the natural environment of the city. With the extensions to "AR Enabling techonogies" and algorithms that we propose for camera tracking, virtual object interaction and NPR illumination coupled under a complete real-time framework for character simulation, we aim provide new dramaturgical notions for Mixed Reality. Such notions could extend further research in MR and develop it as an exciting edutainment medium.

7. Acknowledgements

The authors wish to thank Sebastian Schertenleib for his support with the VHD++ framework and Marlene Arevalo and Nedjma Cadi for their design work. The work presented has been supported by the EU IST FP6 programme, in frame of the EU IST ENACTIVE (http://www.enactivenetwork.org/) project and the EU IST FP5 programme in frame of the EU IST LIFEPLUS (http://www.miralab.unige.ch/subpages/lifeplus) project.

References

[Ans04] Ansari, M. Y., 2004,"Image Effects with DirectX9 Pixel Shaders", ShaderX2:Shader Programming Tips & Tricks with DirectX9, Edited by Engel, W., F., Wordware Publishing Inc., 2004

[ASP05] Archaeological Superintendence of Pompeii, http://www.pompeiisites.org, accessed at 13/06/05

[ART05] ART: Augmented Reality for Therapy, http://mrcas.mpe.ntu.edu.sg/groups/art/, (accessed 10th May 2005)

[ART*05] ARToolKit, http://artoolkit.sourceforge.net/ , (accessed 10th May 2005)

[ASC05] Ascension Technology Corporation, http://www.ascension-tech.com/products/, (accessed 10th May 2005)

[ABB*01] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.,2001, "Recent Advances in Augmented Reality", IEEE Computer Graphics and Applications, November/December 2001

[BP00] Billinghurst, M., Poupyrev, I.,2000, "Shared Space: Mixed Reality Interface for Collaborative Computing", Imagina'2000 Official Guide: Innovation Village Exhibition, 2000, pp. 52

[CMM*03] Cavazza, M., Martin, O., Charles, F., Mead, S. J., Marichal, X., 2003,"Interacting with Virtual Agents in Mixed Reality Interactive Storytelling", 4th International Workshop on Intelligent Virtual Agents, IVA03, 2003

[DK01] Davison, A.J. and Kita, N., "3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain.", IEEE Conference on Computer Vision and Pattern Recognition, 2001

[FHP98] Foxlin, E., Harrington, M. and Pfeifer, G., 1998, "Constellation: A wide-range wireless motion-tracking system for augmented reality and virtual set applications", ACM SIGGRAPH '98, pp. 371-378, 1998

[GHJ*94] Gamma, E., Helm, R., Johnson, R., Vlissides, J., 1994, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1994

[HZ04] Hartley, R.I. & Zisserman, A., 2004, Multiple View Geometry in Computer Vision, Second Edition, Cambridge University Press, 2004

[HMD05] HMD i-Glasses, http://www.i-glassesstore.com/, (accessed 10th May 2005)

[INT05] InterSense, http://www.isense.com/products/prec/ic3/ (accessed 10th May 2005)

[LHM03] Lindt, I., Herbst, I., Maercker, M., 2003, "Interacting within the Mixed Reality Stage", Workshop Proceedings AVIR´03 / Magnenat-Thalmann, Nadia [Eds.], 2003

[Rob05] M. Roberts Motion Control, "Encoded Cranes and Dollies", http://www.mrmoco.com, (accessed 10th May 2005)

[MK94] Milgram, P., Kishino, F., 1994, "A Taxonomy of Mixed Reality Visual Displays", IEICE Trans. Information Systems, vol. E77-D, no. 12, 1994, pp. 1321-1329

[NM00] Nandi, A., Marichal, X., 2000, "Transfiction", Virtual Reality International Conference, Laval May 2000

[PSO*05] Papagiannakis, G., Schertenleib, S., O'Kennedy, B., Poizat, M., Magnenat-Thalmann, N., Stoddart, A., Thalmann, D., 2005, "Mixing Virtual and Real scenes in the site of ancient Pompeii", Computer Animation and Virtual Worlds, p 11-24, Volume 16, Issue 1, February 2005

[PPM*03] Ponder, M., Papagiannakis, G., Molet, T., Magnenat-Thalmann, N., Thalmann, D., 2003, "VHD++ Development Framework: Towards

Extendible, Component Based VR/AR Simulation Engine Featuring Advanced Virtual Character Technologies", IEEE Computer Society Press, CGI Proceedings, pp. 96-104, 2003

[SBT*99] Sannier, G., Balcisoy, S., Magnenat-Thalmann, N., Thalmann, D., 1999 "VHD: A System for Directing Real-Time Virtual Actors", The Visual Computer, Springer, Vol.15, No 7/8, pp.320-329, 1999

[SFC*01] Schwald, B., Figue, J., Chauvineau, E., Vu-Hong, F.,2001, "STARMATE:Using Augmented Reality technology for computer guided maintenance of complex mechanical elements", e2001 Conference, 17-19 October 2001 - Venice – Italy

[SDS*01] Stricker, D., Dähne, P., Seibert, F., Christou, I., Almeida, L., Ioannidis, N.,2001, "Design and Development Issues for ARCHEOGUIDE: An Augmented Reality-based Cultural Heritage On-site Guide", EuroImage ICAV 3D Conference in Augmented Virtual Environments and Three-dimensional Imaging, Mykonos, Greece, 30 May-01 June 2001

[ST02] Strothotte, T. & Schlechtweg, S., 2002, Non-Photorealistic Computer Graphics: Modelling, Rendering and Animation, Morgan Kaufmann Publishers, 2002

[TYK01] Tamura, H., Yamamoto, H., Katayama, A., "Mixed reality: Future dreams seen at the border between real and virtual worlds", Computer Graphics and Applications, vol.21, no.6, pp.64-70. 2001

[TDS*00] Thomas, B., Close, Donoghue, J., Squires, J., De Bondi, P., Morris, M., and Piekarski, W.,2000, "ARQuake: An Outdoor/Indoor Augmented Reality First Person Application", 4th Int'l Symposium on Wearable Computers, pp 139-146, Atlanta, Ga, Oct 2000

[TJN*97] Thomas, G.A., Jin, J., Niblett, T., Urquhart, C., 1997, "A Versatile Camera Position Measurement System For Virtual Reality TV Production.", International Broadcasting Convention, Amsterdam, pp. 284-289, 1997

[UNI05] Unibrain firewire camera, http://www.unibrain.com/home/, (accessed 10th May 2005)

[VLF04] Vachetti, L., Lepetit, V., Fua, P., 2004, "Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking.", Internation Symposium on Mixed and Augmented Reality, Arlington, VA, 2004

[VIC05] Vicon Motion Systems Ltd, http://www.vicon.com. (accessed 10th May 2005)

[WT00] Wohlgemuth, W., Triebfürst, G., "ARVIKA: augmented reality for development, production and service", DARE 2000 on Designing augmented reality environments, 2000, Elsinore, Denmark

III: Simulating Actors and Audiences in Ancient Theaters

Nadia Magnenat-Thalmann, Daniel Thalmann

MIRALab, University of Geneva,
Virtual Reality Laboratory, École Polytechnique Fédérale de Lausanne (EPFL)

Abstract
Ancient theatres were places where a large amount of people from different extractions could gather for important social events such as, but not limited to, representations of comedies or tragedies in which male actors, musicians and dancers wearing masks, in accordance to their specific role and character, would perform on a scene in front of the public. In this case-study we will present the ancient theatre of Aspendos and the odeon of Aphrodisias, located in Turkey near the city of Antalya, virtually restituted in a real-time 3D inhabited environment as Roman buildings of the third century. In order to better understand their use and their feel at Roman times, the choice, and the essential steps that are to be considered, to stage a real-time virtual reenactment of a Roman play will be illustrated: we both focus our attention on the creation of 3D virtual actors capable of performing a selected Roman play and on the creation of a virtual audience that is to emotionally respond to the events that are occurring on stage.

1. Introduction

The theatre of Aspendos (figure 1) is the best preserved roman theatre in Asia Minor, and it is calculated that it was built around 161-180 A.D. during the reign of emperor Marcus Aurelius. In 1078, Pamphylia was claimed by the Seljuk Turks. Certain remains in the theatre of Aspendos and oral sources imply that the theatre was used as a caravanserai (Merchants' Inn) by the Seljuk's. In 1392 ottoman reign sat in and, by the declaration of the Republic of Turkey in 1923, Aspendos took its place within the boarders of the Turkish province of Antalya.

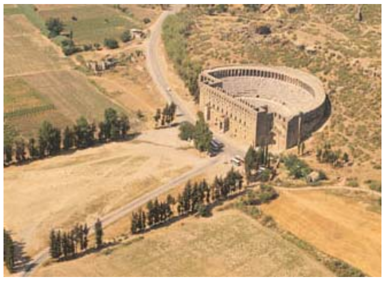


Figure 1. Aerial view of the Aspendos theatre

In the second half of the 19th century, the building caught the attention of scientists. It was the works of researchers such as Lanckoronski [1] Texier [2] and Izenour [3], although occasionally contradicting with each other on certain assumptions, which helped to gradually sweep away the mist of centuries that shadowed Aspendos.

2. Visualization and virtual restoration

Two models of the Aspendos theatre were developed in order to visualize the site in its present state and as it was back in the III century. The main difference between these two versions of the model resides in the elements that were present at the Roman time and that are now no longer present, such as the *Velum*, the roofing of the theatre and all the decorative features of the scene building (general view presented in figure 2).

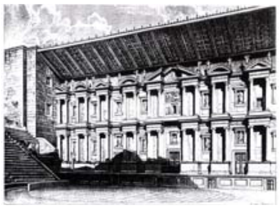


Figure 2. Restitution of the scene building by Lancoronski

3. 3D Environment Modeling

In order to model the Aspendos site virtual environment, accurate topographic data was provided in the form of a 2D elevation map of the area featuring elevation lines every 1 mete. The contours were then used as a base to build a high polygon 3D mesh from which a grayscale elevation map has been extracted to be used to procedurally generate the diffuse textures. In order to meet the needs of a real time simulation, a simplified version of the 3D model of the terrain has been also prepared (figure 3). Figure 4 illustrates a real-time representation of the Aspendos site.

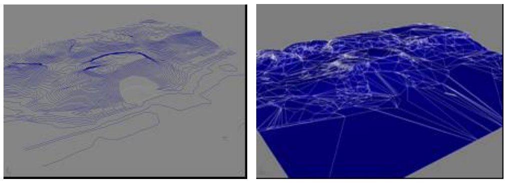


Figure 3. 2D splines extracted from the elevation map positioned in 3D space (left), low polygon optimized mesh of the terrain (right)

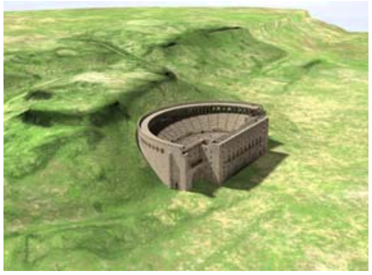


Figure 4. Screenshots from the 3D real time simulation of the Aspendos theatre in its present state

In order to further enhance the visual impact of the real time simulation of the Aspendos site, the addition of diffuse and hard shadows cast by the sun plays a central role. More convincing virtual illumination of the scene is also achieved by applying High Dynamic Range (HDR) Image Based Lighting using virtual light probes. The real-time HDR is used in conjunction with multitextured lightmaps hence allowing the visualization of hard shadows and eliminating at the same time the need of baked textures, thus increasing the sensory level believability.

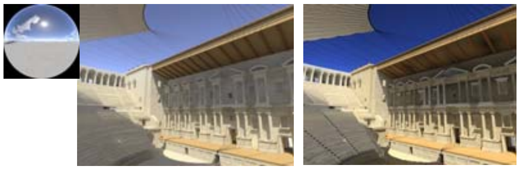


Figure 5. Aspendos theatre rendered using a virtual light probe (left), texture baked 3D model (right) used for the real-time simulation

4. Virtual Actors

In order to better represent the use of the theater at the roman times, the choice to stage a real time virtual reconstruction of a roman play was taken. Therefore virtual embodiments of the actors, meeting the restrictions of a real time simulation, are prepared based on the cultural and historical information about the actors, their outfits and their corresponding masks which are provided by the Yildiz Technical University of Istanbul [3] (figure 6). he current database consists of twelve virtual humans ready for real time animation (presented in figure 7).

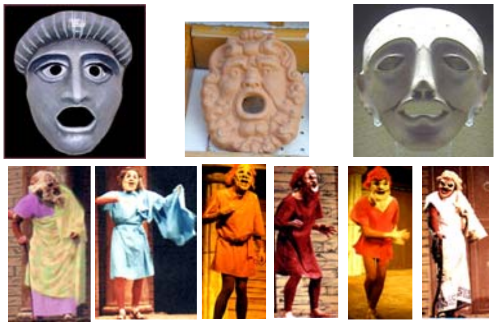


Figure 6. Examples of roman theatre masks (top) and actor's outfits [4] (bottom)

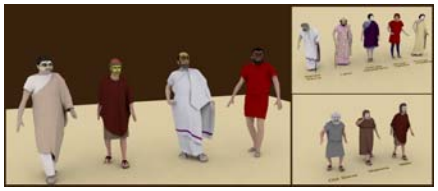


Figure 7. Final database of virtual actors

In order to prepare the models for animation, a template model has been prepared according to an H-Anim LoA skeleton hierarchy. The deformations settings have been parameterized for the 12 designed virtual humans. The captured animation from the models is finally applied to the bodies.

All the developed resources have been combined into two initial applications in order to illustrate the different features of the VR simulations: one based on high fidelity virtual actor simulation and another on large virtual crowd spectator simulation. In the current scenario, the virtual roman actors are reenacting behaviors that involve dialogues, speech with facial animation, body animations, and gestures with object manipulation. The whole crowd, mixed with virtual and real environment, will show emotional behavior with behavior of each actor on the scene.

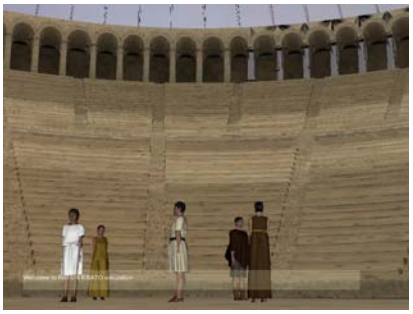


Figure 9: Virtual Characters within the Theater

References

[Lan92] LANCKORONSKI K.: Städte Pamphyliens Und Pisidiens II, Prag, Wien, Leipzig, 1892.

[Tex62] TEXIER C.F.M.: Asie Mineure, Paris, 1862.

[Ize92] IZENOUR G.: Roofed Theaters in Classical Antiquity, Yale University Press, 1992.

[YTU] YTU Project team. February 2004. ERATO Work Package 2 Deliverables, Project No: ICA3-CT-2002-10031, Internal Project Consortium document, Yildiz Technical University.

[Mon] MONAGHAN P.: The Masks of Pseudolus, *Didaskalia* Volume 5 No. 1 - Summer 2001, University of Warwick, edited by Hugh Denard and C. W. Marshall / ISSN 1321-4853

III: Feeling Presence in the Treatment of Social Phobia

Daniel Thalmann, Bruno Herbelin

Virtual Reality Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)

Abstract

The necessity to mixed reality appears recently for patient treatment. Applications of Virtual Reality Exposure have been developed as a tool for mental health therapists for various phobias already. In our case, we applied VRE to the case of social anxiety disorders: social anxiety can be induced by virtual audience and the degree of anxiety experienced is related to the type of virtual audience feedback received by the speaker. We can easily changed the type and the behaviour of the virtual audience (age, sex, emotions, attitude, etc.). We can also analyze the behaviour of the patient using eye tracking. Finally, virtual reality exposure therapy may soon propose specific and unique therapeutic tools.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Three-Dimensional Graphics and Realism]: Presence, Virtual Humans, Mixed Reality, Virtual Reality

1. A Psychotherapeutic Application of VR

Presence is a matter of personnal feelings, a whole body response as if the surrounding events during immersion were real [Sla03]. The original motivation of using VR for mental health was based on this principle. If someone could experience a situation as lived through (i.e. with a sufficient Presence), the psychological impact should be the same than in reality. This short paper will present our work on the virtual reality exposure therapy (VRET) of social anxiety disorder (SAD) relying on such behavioral exercises with VR.

2. The “oral presentation” exposure.

During summer, students ending their formation have to cope with growing stress about their final diploma presentation or job interviews. However, only a few feel so anxious that they finally ask for the support of a psychotherapist. We experimented, with two young patients who had a pathological social phobia a public-speaking VRE where they could practice oral presentation in front of a virtual assembly. The immersive virtual environment figures a typical public speaking situation with people sitting and listening (fig.1). The speaker stays in front of them, standing or sitting on a chair (depending on the immersion setup). While doing his/her presentation, he/she can observe the assembly listening with more or less attention and showing typical attitudes (crossing their legs or arms, jawing, touching their hair, etc.). Our experiments with these patients were fruitful for several reasons. First, we could see positive results in their rehabilitation after only four sessions (e.g. one finally succeeded in his diploma oral presentation). And second, we could make some important correlations between their perception of the virtual assembly and their fears in reality.

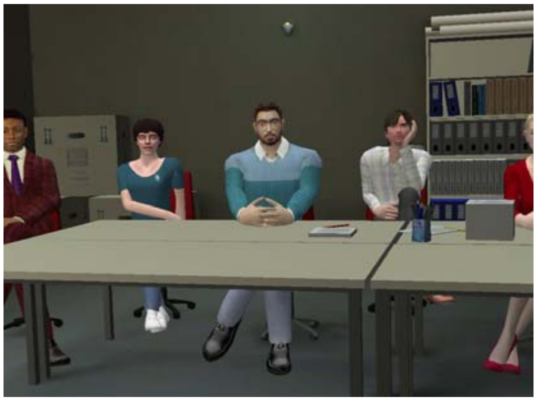


Figure 1. Virtual scene for public-speaking exposure therapy.

3. Analysis.

In the same way North et al. [NSM02] conclude that “VRE may be used as an effective treatment method for reducing self-reported anxiety” for social phobia, we could observe a very positive impact of these sessions on the successfulness of the therapy. Similarly to Pertaub et al. [PSB01], our therapist noticed, both in the patients’ behaviors and during classical post-exposure therapeutic sessions, that the subjects were very sensitive to the behaviors of the virtual audience. They considered the presence of the virtual assembly as real enough to generate stress (i.e., they had similar feelings as in reality: difficulty speaking, hands shaking, gaze avoidance).

It follows that the behaviors observed in VRE can be analyzed under these conditions and corrected during exposures as efficiently as in-vivo. The main advantage with the VRE is the possibility to track, record and analyze the patent's behavior during the simulation with computerized tools providing more information than the usual recordings on video tapes. Based on this idea and referring to the observations made by Horley et al. [HWGG03], the gaze behavior of social phobic patients in VRE should be characterized by 'eye to eye' avoidance. For this reason, we experimented with eye tracking technology to have the possibility to analyze the patient's gaze focus –when looking at a person and more specifically looking in the eyes. The therapist started to use this tool to keep records of patients' performance and to observe their improvements. We could observe that the device offered sufficient precision to distinguish not only who the user was speaking to but also which part of the body he was looking at --although distinguishing the part of the face (e.g. the eyes) is only possible if the user gets closer to the virtual character.

The next step will be to integrate the virtual assembly directly into the real physical place with AR. By focusing on social presence, and hopefully removing the immersion bias, we may reach even more efficient results.

4. Conclusion.

The mental health community is today quite confident on the positive therapeutic impact of VRE since it has been successful for now ten years. Our simulation of virtual assembly also proved to be believable enough for the therapy of social anxiety disorders.

However, we consider that VRE should not be limited to a “virtual reality on demand”, like a song on a CD or a movie on a DVD, but as the potential behavioral assesment tools for clinical evaluation and self-rating reference for exposure and therapeutic exercices.

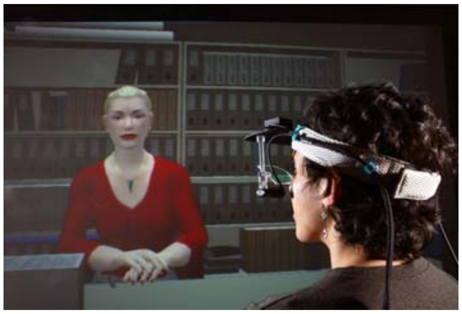


Figure 2. Eye tracking during VRE for gaze behavior analysis and avoidance correction.

References

[PSB01] Pertaub, D. P., Slater, M., & Barker, C. (2001). An experiment on fear of public speaking in virtual reality. *Studies in Health Technology and Informatics, 81,* 372–8.

[Sla03] Slater, Mel. (2003) A note on Presence terminology. *Presence-Connect (on-line)*, jan 2003.

[NSM02] North, M.M., Schoeneman, C.M. and Mathis, J.R. (2002). Virtual reality therapy: case study of fear of public speaking, in *Medicine Meets Virtual Reality* 02/10, pp.18-20.

[HWGG03] Horley, H., Williams, L.M., Gonsalvez, C. and Gordon,E. (2003). Social phobic do not see eye to eye: A visual scanpath study of emotional expression processing, *Journal of Anxiety Disorders,* vol. 17, pp.33-44.

III: A mixed-reality system for surveillance and security

Mario Gutierrez, Renaud Ott, Sylvain Cardin,

Virtual Reality Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)

Edited by Frédéric Vexo

Abstract

We present a system that exploits advanced Mixed Reality technologies to create a surveillance and security system. Surveillance cameras are carried by a mini Blimp which is tele-operated using an innovative Virtual Reality interface with haptic feedback. An interactive control room (CAVE) receives multiple video streams from airborne and fixed cameras. Eye tracking technology allows for turning the user's gaze into the main interaction mechanism; the user in charge can examine, zoom and select specific views by looking at them. Video streams selected at the control room can be redirected to agents equipped with a PDA. On-field agents can examine the video sent by the control center and locate the actual position of the airborne cameras in a GPS-driven map. The PDA interface reacts to the user's gestures. A tilt sensor recognizes the position in which the PDA is held and adapts the interface accordingly.

Categories and Subject Descriptors (according to ACM CCS):
H.5.2 [User Interfaces]: Haptics I/O
J.7 [Computer in Other Systems]: Command and Control

1. A mixed-reality system for surveillance and security

As a case study we present a mixed-reality system that exploits advanced data acquisition and rendering technologies to create a surveillance and security system. Surveillance cameras are carried by a mini Blimp which is tele-operated using an innovative mixed-reality interface with haptic feedback. An interactive control room (CAVE) receives multiple video streams from airborne and fixed cameras. Eye tracking technology allows for turning the user's gaze into the main interaction mechanism; the user in charge can examine, zoom and select specific views by looking at them. Video streams selected at the control room can be redirected to agents equipped with a PDA. On-field agents can examine the video sent by the control center and locate the actual position of the airborne cameras in a GPS-driven map. The PDA interface reacts to the user's gestures. A tilt sensor recognizes the position in which the PDA is held and adapts the interface accordingly. The prototype we present shows the added value of a mixed-reality application and opens up several research directions in the areas of tele-operation, Multimodal Interfaces, etc.

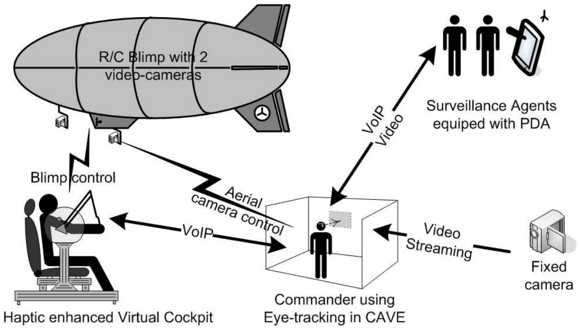

R/C Blimp with 2 video-cameras
Surveillance Agents equiped with PDA
VoIP Video
Blimp control
Aerial camera control
VoIP
Video Streaming
Fixed camera
Haptic enhanced Virtual Cockpit
Commander using Eye-tracking in CAVE

Figure 1: *The overall system architecture.*

Figure 1 shows the different modules that compose the surveillance and security system. We can distinguish three main parts :

- Control of the aerial device (the Blimp) supporting the video cameras. This task is done by a single pilot seating on Haptic Workstation™ inside a distant and closed environment. The pilot can control the blimp as if he were inside it.

- On-field agents : They are equipped with handheld devices in order to receive precise orders including multimedia content (text, images, sound).
- Coordinating on-field agents and blimp pilot. A commander communicates with the pilot and every agent and gives them spoken orders. He has also a real-time view of all the cameras (mobile and fixed), and can also send multimedia content to on-field agents.

R/C Blimp

The blimp shown in figure 2, is a low-cost Unmanned Aerial Vehicle (UAV) that we use in our teleoperation research.

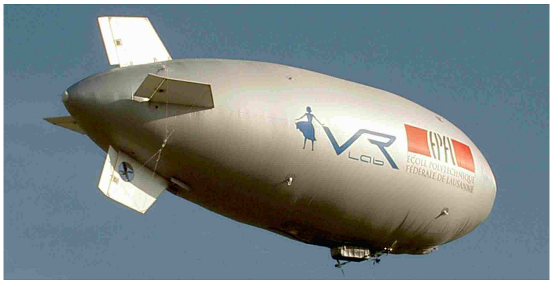


Figure 2: *R/C blimp.*

The *R/C Blimp* is composed by a $6,75m$ long and $2,20m$ diameter envelope that is filled with $11m^3$ of Helium gas (He). The total weight including its standard flight equipment is $9kg$, there is around $2kg$ of maximum payload for the cameras and the video transmission system. Below, there is a gondola containing the electronics part, the power supply ($6600mAh$ allowing $1h$ at half-speed) and supporting the two electric motors. Each have $1,5kg$ of power, allowing the blimp to fly at $35km/h$ when there is no wind. The range of the transmission of the radio controller is $1.5km$, but it can be extended with repeaters.

Figure 3 describes the computers used to control and gather information from the blimp. Communications are near real-time because everything is done in hardware: the video is delayed for less than $50ms$, and the control of the blimp via the servo controller is in real-time. In the next subsection we will describe the *Virtual Cockpit* system used to control this Blimp.

Virtual Cockpit

The *R/C blimp* is remotely controlled from a *Virtual Cockpit*. The interface must be **precise** (for a fine control), **instructive** (to give location information) and **intuitive** (to avoid manipulation errors). The visual part of the blimp is rendered to the pilot via and HMD. In order to have a virtual camera that moves according to the head movements, we have used an orientation tracker, the InertiaCube3[Int]. We use a realtime OpenGL viewer created in our laboratory. It allows for rendering stereo 3D models and videos at the same time by mapping streams on a polygon. Figure 4 shows the representation of the blimp within the mixed-reality environment. The user sits inside a 3D gondola made of glasses. Behind the glasses, we display a video stream coming from a camera mounted on the blimp. The GPS information coming from the blimp is overlayed on the window and represented as a 2D map indicating the blimp's location. Heading, altitude, and speed (horizontal and vertical) are displayed as well.

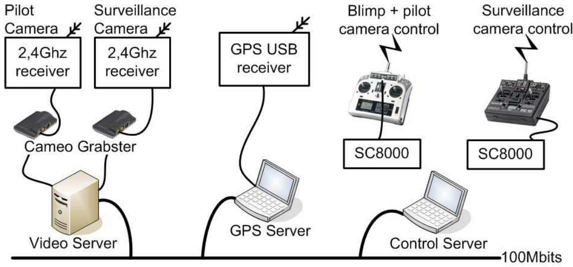


Figure 3: *The Blimp communications system: three pc connected on a 100Mbits network for receive video and GPS and controlling servos.*

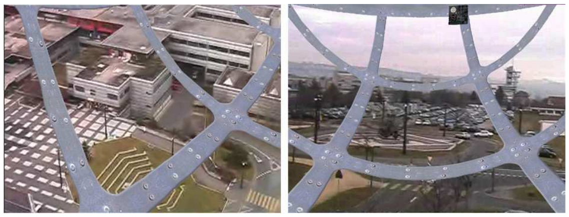


Figure 4: *The mixed-reality Blimp cockpit.*

The actual piloting of the blimp is done through a Haptic Workstation™. In [OGTV05b], we have shown that teleoperation of a vehicle using haptic devices is more efficient when having a gesture interface. Interaction within the mixed-reality cockpit is as follows: right hand is used to control the aerodynamic stabilizer (by moving the hand forward/backward and left/right), whereas the left hand is used to control the engine power. When there is no power, a left/right hand movement doesn't move the stabilizers but controls the rear engine allowing the pilot to do fast aboutturn, useful in urban environments. One disadvantage of the Haptic Workstation™ is that it is uncomfortable during long sessions, because the arm must be kept outstretched while supporting the weight of the exoskeleton. In [OGTV05a], we presented a software for improving the comfort when using an haptic device by creating a zero-gravity illusion. We have used such system in order to compensate the weight of both the user arms and the exoskeleton. Moreover, the force feedback constraints the user hands to the neutral position, i.e. no power and straight ahead. When the user wants to move a hand, the force feedback intensity increases. The pilot can thus feel how fast he is going by evaluating the force applied on his hands. In the next subsection, we will study the *Surveillance Control Room* where orders to the pilot are given.

Surveillance Control Room

In order to place the supervisor in the best disposition for taking decisions, we focused on designing an ergonomic interface using Virtual Reality devices. Our system displays several video streams in a CAVE, and allows the supervisor to select and send visual information to the *On-field Agents*.

Figure 5 shows the system in charge of controlling the CAVE.

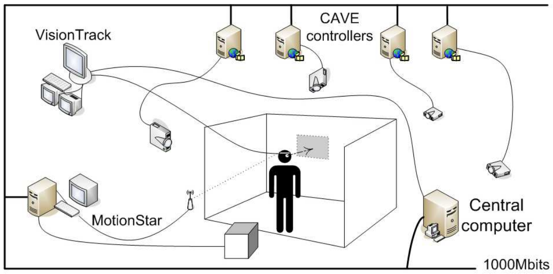


Figure 5: *The control room system.*

The CAVE offers the possibility to display multiple video streams on four projection screens (see figure 6), providing full immersion of the user into the scene.

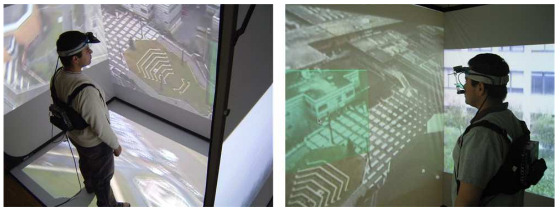


Figure 6: *Four sided CAVE.*

A joystick allows the supervisor to move mobile cameras to get the most appropriate point of view (see figure 7). Orientation is controlled by the pad and two buttons are used to zoom in and out. By pressing a single button, the user validates his eye selection and sends it to the *On-field Agents.* We chose to send a single relevant picture rather than directly the video stream viewed by the supervisor for readability reasons and network resources management. Passing the right information at the right moment is crucial. Since our camera is aerial, it gets a field of view covering a large area. In contrast, agents on the ground get a local view of the situation.

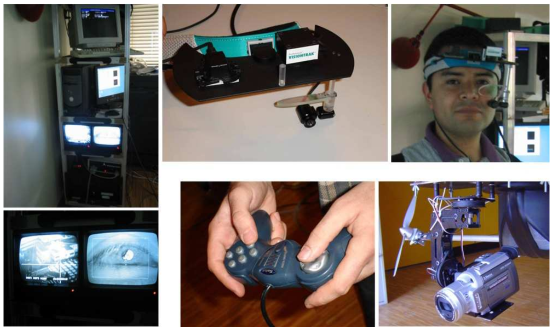


Figure 7: *Top and left: The Eye-tracking system. Bottom-right: The mounted rotating camera and the joystick controller.*

In order to select part of the image, the eye tracking system described in the part of the tutorial titled: Hardware for Mixed Realities in Inhabited Worlds is used to compute the gaze direction and determine the gaze target on the CAVE. This system provides the supervisor with powerful tools to send visual and vocal information to *On-field Agents*.

On-field Agent Equipment

We have designed a handheld interface that minimizes the use of buttons and menus and responds to intuitive gestures from the user. Our handheld communication equipment (PDA) provides a dynamic interface that reacts to the way it is held by the agent.

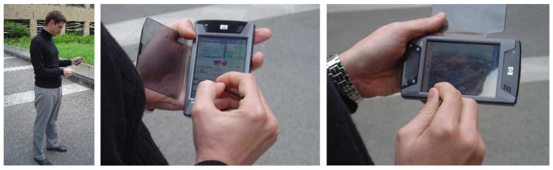


Figure 8: *On-field agent communication equipment.*

The system is based on commercial PDA devices using the PocketPC operating system with Wireless LAN capabilities. By means of a tilt sensor [ECE], the PDA detects whether it is positioned horizontally or vertically and presents different information. Live video streaming selected at the control room is displayed on the handheld, as well as map of the site enhanced with GPS information coming from the aerial monitoring system (R/C blimp).

When held vertically, see figure 8, the interface shows a reduced image of the live video stream selected at the control room as well as a map of the surveyed site. GPS information is used to point out on the map the current position of the blimp. When held horizontally (figure 8), the interface shows a higher resolution view of the live video stream. This way, *On-field Agents* can better appreciate specific details while maintaining communication with the control room. Communication with the commander at the control room is done through Voice over IP.

Conclusions on the Case Study

We have described a full surveillance and security system based on advanced Virtual Reality technologies. Our system can be applied to the surveillance of public sites such as stadiums, universities, parks, etc. Our contribution focuses on the successful application of state of the art Virtual Reality and multimedia technologies to create a in a complex mixed-reality system with multiple participants. Our main goal was to maximize interaction and communication between the personnel implied in the tasks. A secondary goal was reducing the amount of people needed for operative tasks that were not directly related to the actual surveillance and security activities. VR technologies eliminate the need for secondary/auxiliary operators and provide intuitive and more natural interfaces. We use haptic interfaces for teleoperating a flying vehicle. Our results are encouraging in the sense that the haptic interface works as well as a classical control console. Future work consist on taking full advantage of the new dimension provided by haptic feedback for conveying additional information to the pilot and ease the piloting task. In terms of information analysis, we have obtained satisfactory results with the use of eye tracking technology within a CAVE system. We proposed an innovative interface for picking-up zones of interest from live video stream. The system is complemented with an efficient multimodal communication device based on a PDA. This solves a common demand of security agents who require more than just voice communication with the control room. Our system enables the commander at the control room to send live video stream of the zones of interest that require a special attention. This provides *On-field Agents* with valuable information and facilitates their task. This kind of mixed-reality applications fit into the context of current initiatives for security and enhanced surveillance systems.

References

[ECE] ECER TECHNOLOGY: TiltControl Device http://www.ecertech.com. 3

[Int] INTERSENSE: http://www.isense.com/. 2

[OGTV05a] OTT R., GUTIERREZ M., THALMANN D., VEXO F.: Improving user comfort in haptic virtual environments trough gravity compensation. In *Proceedings of WorldHaptics'05* (2005), pp. 401–409. 2

[OGTV05b] OTT R., GUTIERREZ M., THALMANN D., VEXO F.: Vr haptic interfaces for teleoperation : an evaluation study. In *Proceedings of the IEEE Intelligent Vehicles Symposium, IV'05, (to appear)* (2005). 2