

Interacting with Virtual Agents in Mixed Reality Interactive Storytelling

Marc Cavazza¹, Olivier Martin², Fred Charles¹, Steven J. Mead¹
and Xavier Marichal³

(1) School of Computing and Mathematics, University of Teesside,
Borough Road, Middlesbrough, TS1 3BA, United Kingdom.

(2) Laboratoire de Télécommunications et Télédetection,
Université catholique de Louvain, 2 place du Levant,
1348 Louvain-la-Neuve, Belgium.

(3) Alterface, 10 Avenue Alexander Fleming, 1348 Louvain-la-Neuve, Belgium.
{m.o.cavazza@tees.ac.uk, martin@tele.ucl.ac.be,
f.charles@tees.ac.uk, xavier.marichal@alterface.com,
steven.j.mead@tees.ac.uk}

1 Introduction

User interaction with virtual agents generally takes place in virtual environments in which there is clear separation between the virtual actors and the user, due to the fact that in most cases, the user is in some way external to the virtual world. In Mixed-Reality Interactive Storytelling, the user's video image is captured in real time and inserted into a virtual world populated by autonomous synthetic actors with which the user interacts. The user in turn watches the composite world projected on a large screen, following a "magic mirror" metaphor. This context leads to re-investigating the techniques by which the user interacts with virtual actors, as well as exploring specific research problems. In this paper, we discuss some specificities of user interaction with virtual actors in Mixed Reality Interactive Storytelling. After a brief introduction to our system's architecture and the example scenario supporting our experiments, we describe various techniques supporting multi-modal interaction with virtual actors.

2 System Architecture

Our Mixed Reality system is based on a "magic mirror" model (Figure 1), in which the user's image is captured in real time by a video camera, extracted from his/her background and mixed with a 3D graphic model of a virtual stage including the synthetic characters taking part in the story (Figure 2). The resulting image is projected on a large screen facing the user, who sees his own image embedded in the virtual stage with the synthetic actors. The graphic component of the Mixed Reality world is based on a game engine, Unreal Tournament 2003TM. This engine not only performs graphic rendering and character animation but incorporates a new version of our previously described storytelling engine [2].

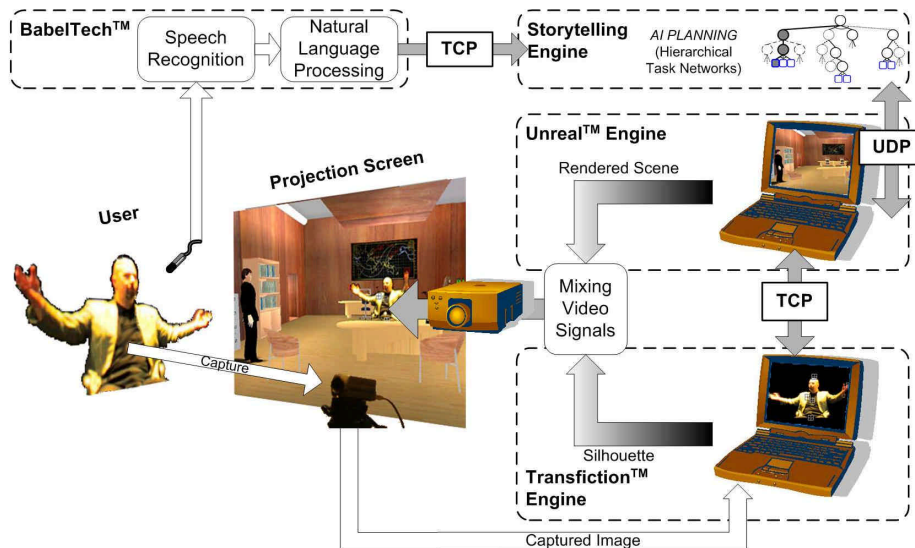


Fig. 1. System Architecture.

A single 2D camera facing the user analyses the image in real-time by segmenting the user's contours [5]. The objective behind segmentation is twofold. Firstly, it extracts the image silhouette of the user in order to be able to inject it into the virtual setting on the projection screen. Secondly, it can recognise user behaviour, including symbolic gestures, in real-time, hence supporting a new interaction channel.

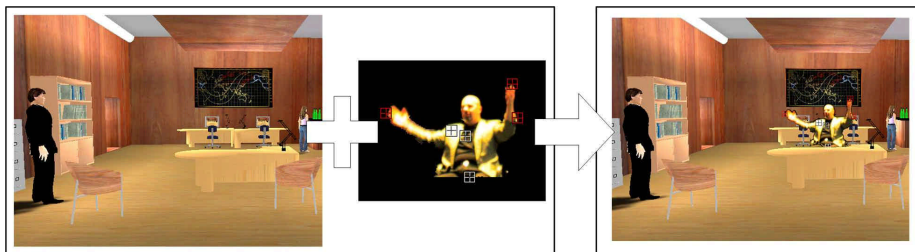


Fig. 2. Creating Mixed Reality Scenes.

The storytelling scenario supporting our experiments is based on a James Bond adventure, in which the user is actually playing the role of the villain. James Bond stories have narrative properties that make them good candidates for interactive storytelling experiments: for this reason, they have been used as a supporting example in the foundational work of Roland Barthes in contemporary narratology [1]. Besides, their strong reliance on narrative stereotypes facilitates narrative control and the understanding of the role that the user is allowed to play. The basic storyline represents the early encounter between Bond and the villain (let us call him the Professor). The objective for Bond is to acquire some essential information, which he can find by searching the Professor's office, obtain from the Professor's assistant or even, under certain conditions, (deception or threat) by the Professor himself. The actions of the user (acting as the Professor) are going to interfere with Bond's plan, altering the unfolding of the scene.

Like many other interactive storytelling systems, our prototype is based on a cast of virtual actors controlled by real-time planning systems, whose contents in terms of plan formalise each actor's role in the baseline plot [2]. It can be noted that the main character in the story is actually not the user but a virtual actor impersonating the James Bond character. To that extent, its role and, from a technical perspective, the unfolding of its corresponding plan, will be the main driver for the interactive narrative.

3 User Interaction with the Virtual Actors

The fact that the user visually takes part in the story presentation obviously affects the modes of user intervention: these will have to take the form of traditional interaction between characters. In other words, unlike in some of the systems we previously developed, the user will have to *act*. This means that the actions by which he may interfere with the story should have a visual presentation that blends into it. In other words, the mechanisms of his normal acting (gestures, speech) should serve as a natural basis for his intervention in the storyline. These mechanisms comprise physical interaction (the direct contact between the user and virtual world entities), symbolic gestures, and speech (the latter two combining through various forms of multi-modal interaction).

Physical interaction consists in all forms of contact between the user (or more precisely the user's embodiment through his video avatar) and virtual actors. This interaction is mediated by the low-level mechanisms managing interaction between actors in the Unreal Tournament engine. The user's position is associated with an empty bounding box that can generate various kinds of interaction information with the environment's objects and the virtual actors. Physical interaction is based on the use of a similar co-ordinate system for both the virtual world and the video information captured from the real setting. As a consequence, the user's bounding box follows the user's movements in the real world. For instance, when the user moves towards a virtual actor (which results in his bounding box colliding with that of the agent), several forms of interactions become active (such as attracting attention, shaking the agent's hand, slapping or punching the agent).

The exact form of physical interaction is triggered by the simultaneous recognition of a user's gesture and a collision between bounding boxes. This will in turn determine the action performed on a virtual actor and affect the storyline accordingly. For instance, depending on the gesture recognised (a handshake or a slap), the corresponding agent reaction is displayed (the appropriate animation being played) and the consequence of this reaction is used to update the agent's plan.

The same mechanism can be used to interact with the virtual world objects. However, in a Mixed Reality context the interactions between the user and virtual world objects are limited by traditional problems of transition between the real and the virtual world, such as passing objects from the real to the virtual world, for which they exist no generic solutions.

User gestures are recognised through a rule based system that identifies gestures from a gesture library, using data from image segmentation that provides in real time the position of user's extremities. One essential aspect of the interaction is that the

system is actually tracking symbolic gestures corresponding to narrative functions, such as greetings, threatening (or responding to a threat, such as putting his hands up), offering, calling, dismissing, etc. However, gestures are mostly used in conjunction with speech recognition. The accurate recognition of a user intervention is based on i) the current stage of the plot and ii) a multi-modal interpretation. For instance, the fact that the user stands up from a seated position could be interpreted as a greeting at the very beginning of the scene, while it would be interpreted as terminating the interview should it take place at a later stage.

The speech recognition component is based on the Ear SDK from BabelTech™, which can be used in various modes including multi keyword spotting in an open-loop mode. Speech and gesture information is combined using temporal information as classically described for multi-modal utterances. The objective of the multi-modal parser is to identify speech acts that can be interpreted by the virtual actors, such as e.g., greetings, questions, information provision, or threats. These speech acts can then be mapped to the actor's plan to modify its progression [3] [4].

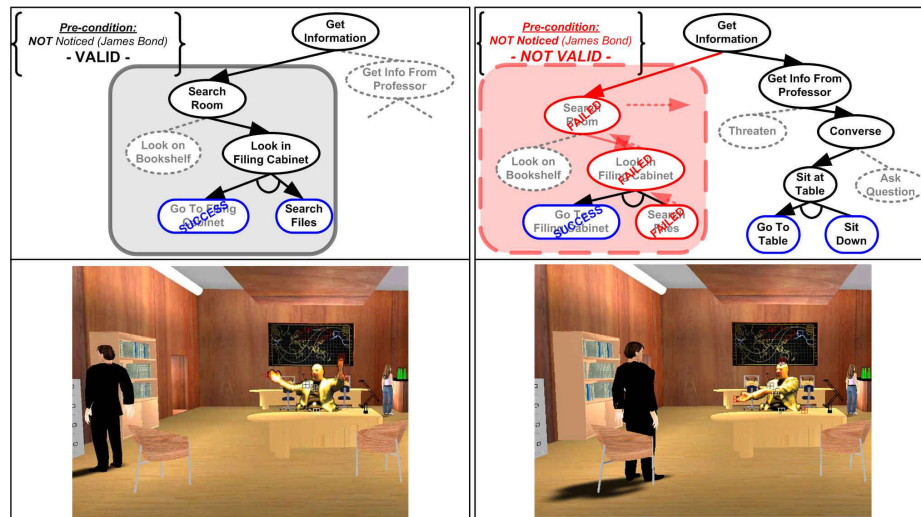


Fig. 3. An Example of User Intervention.

The user's greetings force a change of plans in the main character.

To illustrate briefly this implementation of mixed reality interactive storytelling, we can consider the partial example presented on Figure 3. At this early stage of the plot, Bond has entered the Professor's office and has started searching for documents in the filing cabinet, thinking the room was empty. From a planning perspective, his current goal is to acquire information and one of the sub-goals is to search the filing cabinet for relevant files. However, a condition for such search is to remain unnoticed. When the user greets him, Bond becomes aware of the professor's presence and has to direct himself towards him, abandoning his current actions. From an acting perspective, greetings consist in a combination of a spoken utterance (such as "welcome, Mr. Bond!") and a greeting gesture. The speech acts also triggers a corresponding response in terms of plan (a situated action), by which the Bond character joins the professor at his desk. After which the normal plan can be resumed,

whose further course will depend on the various actions taken by the user, as well as the intervention of other characters in the story.

4 Conclusions

Mixed Reality offers a new context for Interactive Storytelling, which puts the user in a double actor-spectator position, through the “magic mirror” metaphor, which provides an inverted third-person mode of participation. This is a significant departure from other paradigms of user involvement, such as pure spectator (with the ability to influence the story) [2] or an actor immersed in first-person mode following a “Holodeck™” paradigm [6]. While the practical implications of this form of user involvement are yet to be explored, the same context brings new perspectives for user interaction as well, with an emphasis on multimodal interaction. Our current work is dedicated to improving the image processing components of the system (mainly video fusion and partial resolution of occlusion problems) and to scaling-up the multi-modal interaction component, in particular its speech recognition module.

Acknowledgements

Olivier Martin is funded through a FIRST Europe Fellowship provided by the Region Wallonne.

References

1. R. Barthes, Introduction à l'Analyse Structurale des Récits (in French). Communications, 8, pp.1-27, 1966.
2. M. Cavazza, F. Charles, and S.J. Mead, Character-based Interactive Storytelling, IEEE Intelligent Systems, special issue on AI in Interactive Entertainment, pp. 17-24, 2002.
3. M. Cavazza, F. Charles, and S.J. Mead, Under The Influence: Using Natural Language in Interactive Storytelling, 1st International Workshop on Entertainment Computing, IFIP Conference Proceedings, 240, Kluwer, pp. 3-11, 2002.
4. S.J. Mead, M. Cavazza, and F. Charles, “Influential Words: Natural Language in Interactive Storytelling”, 10th International Conference on Human-Computer Interaction, Crete, Greece, 2003, Vol 2., pp.741-745.
5. X. Marichal, and T. Umeda, “Real-Time Segmentation of Video Objects for Mixed-Reality Interactive Applications”, Proceedings of the "SPIE's Visual Communications and Image Processing" (VCIP 2003) International Conference, Lugano, Switzerland, 2003.
6. W. Swartout, R. Hill, J. Gratch, W.L. Johnson, C. Kyriakakis, C. LaBore, R. Lindheim, S. Marsella, D. Miraglia, B. Moore, J. Morie, J. Rickel, M. Thiebaut, L. Tuch, R. Whitney, and J. Douglas, “Toward the Holodeck: Integrating Graphics, Sound, Character and Story”, in Proceedings of the Autonomous Agents 2001 Conference, 2001.