Multilingual Virtual City Guides

Karina Rodriguez Echavarria, Michel Genereux, David B. Arnold
Faculty of Management and Information Sciences
University of Brighton

Andrew M. Day, John R.W. Glauert
School of Computing Science
University of East Anglia

Abstract

In this paper the potential of combining low cost systems with technologies for modelling and rendering populated urban scenes and multilingual interactive avatars is presented. As an example, a medieval European city was recreated, such that within this virtual environment, visitors can see for themselves how buildings look and hear from a multilingual interactive guide the history of the place. The selected route along with the visitor profile, such as their language and age, influence the way in which the information related to the city is presented. This provides more natural and engaging interactions within the virtual environments, which may be further improved by using affective and stylized speech output Thus, it enhances the learning and entertainment experience of the users.

Keywords: Urban environments, Scene Assembly, Avatars, Natural Language Generation

1. INTRODUCTION

European towns and cities have a huge wealth of historically significant and culturally important material in the form of the physical environment (i.e. buildings and open spaces) and intangible heritage (e.g. events). These embody local cultures and traditions through time and in turn may shape the developing culture of current societies which occupy those spaces. Therefore, their re-creation can promote and provide a better understanding of European culture.

Virtual Environments offer the potential of recreating towns and cities for visitors to see for themselves how buildings once looked and to hear from virtual guides the history of the place. As processor speeds and memory sizes increase and the cost of hardware falls, the possible results become more impressive. Nevertheless, major difficulties still remain for widely adopting these technologies. High on the list are the complexity of building the virtual environments [1] and the challenge of providing natural and enjoyable ways of interaction.

Traditionally 3D modelling has been labour intensive and time consuming. The traditional approach to build 3D environments has been to create highly detailed models with modelling toolkits such as 3Dstudio Max or Maya. These models are then exported to an all-purpose rendering engine to interactively explore the virtually reconstructed city. Some of the major problems with this approach are:

- Knowledge of the application domain is not exploited to simplify the visitor interactions.
- The models are not optimised for real time rendering of the complex scene.
- Expert knowledge of the technologies is required to create models that are efficient at run time, effectively excluding many content creators who may have the knowledge to create more engaging experiences.

In order to overcome these problems, more specialized modelling tools that exploit domain knowledge have been created. The CHARISMATIC project [2] created a toolkit to support the construction of large urban scenes with specially designed tools and techniques for modelling buildings, avatars, trees, and generic 3D objects.

In addition, providing multimodal and natural language interaction including in the longer term emotion and style, within these virtual environments is essential for enhancing the learning and entertainment experience of the users. To achieve this, multilingualism is required to allow as many users as possible to benefit from these experiences [3]. These issues involve tackling traditional problems in computational linguistics and speech processing, but also more recent research areas such as multimodality, context-sensitivity and the specific modelling of cultural heritage sites visitors. These issues have been examined in projects such as the M-PIRO project [4] which automatically generated multilingual descriptions of museum exhibits according to users' backgrounds, ages, and previous interaction with the system. Another example, NICE, presents a conversational avatar, discussing the life and work of Hans Christian Andersen [5]..

This paper presents the potential of combining low cost systems with technologies for modelling and rendering populated urban scenes and multilingual interactive avatars. The use of avatars guides as a means of exploring the virtual environment provides the visitor with a more natural and engaging interaction. As an example, a medieval European town was recreated. Within this virtual environment, a virtual avatar guides the visitors through the routes of the city which suit their interest. The selected route along with the visitor profile, such as their language and age, influence the way in which the information related to the city is presented by the guide. The results achieved from this work are presented in the following section, followed by conclusions and further research directions.

2. WOLBENBÜTTEL VIRTUAL ENVIRONMENT

The environment produced in this work recreates a medieval European town from Lower Saxony (Germany) called Wolfenbüttel where most buildings date from the seventeen century. Wolfenbüttel became the residence for the dukes of Brunswick in 1432 and in the following three centuries the town was an important centre of the arts. The virtual environment developed recreates the main buildings of the town, such as the duke's palace, the library, the armoury, as well as other areas of interest. The virtual environment offers the possibility of exploring the town and interacting with a multilingual virtual guide who provides information regarding these buildings and of the town in general in a more enjoyable way. The languages included in this environment are English and French.

The approach used to develop this virtual environment was the following:
1. Modelling the urban scene and important buildings
2. Populating the scene with avatars
3. Creating a multilingual interactive avatar
4. Developing the multimodal and multilingual language generation module in order to drive the interaction of the visitor with the virtual environment.

The following subsections will describe in more detail each of these steps.

2.1 Modelling the urban scene and important buildings

The town scene consisted of a small (500 m2) region with important buildings and a much larger surrounding region of generic houses. A large proportion of the scene was constructed in a relatively short time by using the CHARISMATIC toolkit to rapidly create the generic houses and automatically position them along the roads. The open-source scengraph OpenSG was used during this process. The important buildings of the town were then modelled using 3DStudio Max and included in the scene (see figure 1). Finally, parametric trees were added to the scene for realism.

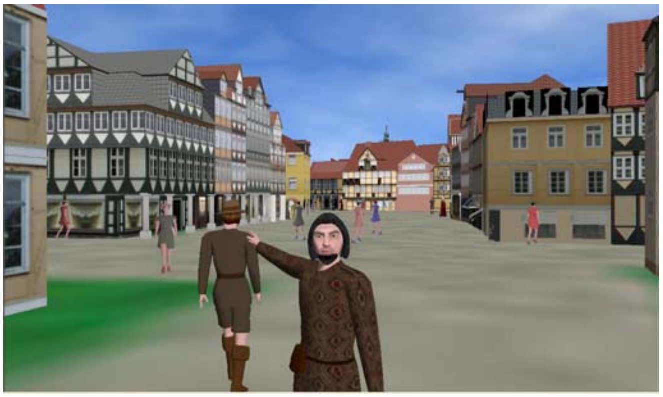


Figure 1: Example of buildings modelled in Wolfenbüttel 's virtual environment

2.1.1 Improving rendering of scenes

In order to maintain interactive frame rates (25fps and above) and due to the large scale of the scene created, it was required to use methods of reducing the number of geometric features to be rendered in any given animation frame. The techniques used included subdivision surface work from Braunschweig University [6] and Occluder Culling [7]. This technique follows the principle that in urban environments, every building in the scene potentially occludes other buildings and avatars. By testing buildings and avatars against foreground buildings before they are rendered it is possible to send a smaller number of objects for rendering to the graphics card.

2.2 Populating the scene

In order to bring life to the scene, a number of virtual people were added around the town. There are several situations where the fact that the scene contains moving people (both full geometry and crowd impostors) brings the danger that they may interact with the terrain and buildings in an uncontrolled or undesirable manner. However, Wolfenbüttel is quite a flat scene, so it was possible to ignore the problem of generating realistic animation of people walking on sloping or uneven terrain.

The problem of potential collision detection between the moving avatars and the buildings remains however. Due to the large number of avatars in the crowds, full implementation of collision detection was deliberately avoided as it was computationally too expensive. Instead, a grid-based collision-avoidance method was used, where each metre-square is either accessible to the avatars or not. This works well for open spaces, but it prevents the avatars from navigating freely in some of the narrow spaces in the streets of the town. To stop the avatars becoming trapped, the narrow passages were manually closed. This could have been avoided with a full collision detection system, or more sophisticated crowd behaviour/collision avoidance rules.

2.3 Creating interactive avatar

In order to make the visitor interaction with the virtual environment as natural and engaging as possible, speaking avatars were created. The aim was to have a multilingual interactive avatar to guide the visitor of the virtual environment and to respond to predefined questions related to the building and events in the town. These responses were tailored depending upon knowledge about the visitor and previous requests. To achieve this, several issues had to be addressed: mouth shapes, gestures as well as speech and its lip synchronisation. These issues will be presented in the following subsections.

2.3.1 Morph target creation

The parameters derived from the speech recognition database were used to activate primitive morph targets on the avatar which were blended together to achieve the correct mouth shapes. The primitive shapes were taken from the Facial Action Coding System (FACS) [8]. This is a minimum set of mouth shapes based on muscle movement in the face (see figure 3). At present, only five parameters were implemented: one for jaw movement, one for tongue movement up and down, and three for the mouth.

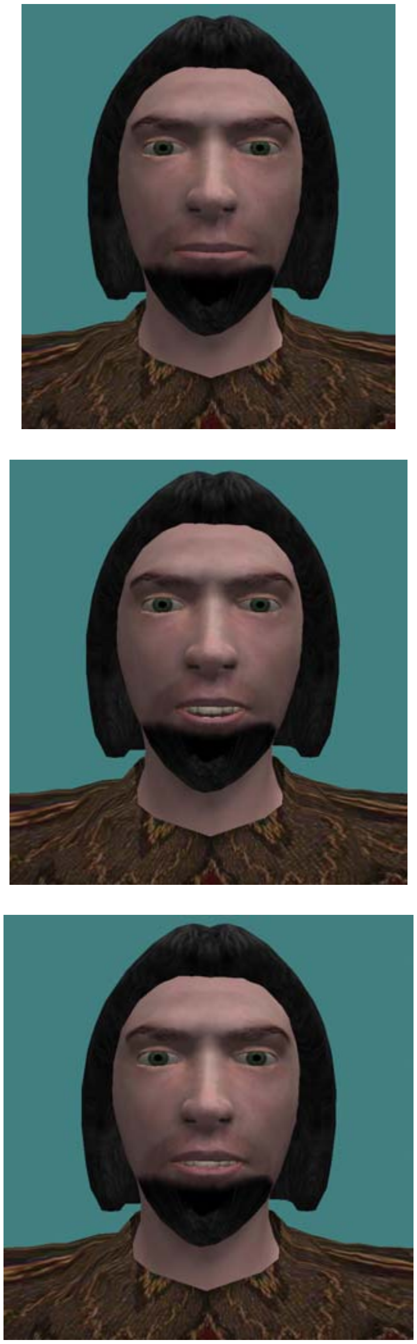


Figure 3: Example of primitive mouth shapes

The morph targets were created using a toolkit developed at UEA that enables interactive selection and weighting of vertices on the avatar face mesh, with facilities to apply transformations on these vertices including translation, rotation, and growth (expansion along vertex normals) to create the final morph target. The morph targets were named according to the face parameters, uploaded into the avatar, then blended and applied on a frame-by-frame basis.

To achieve correct shading of the deformed mesh, vertex normals for those vertices affected by the morph targets were calculated in real time. From work carried out the UEA [9], evaluations of the contribution made by shadows around the mouth and surrounding area as it deforms have been shown to contribute significantly to recognition by users. Vertex normals for the unaffected vertices of the avatar mesh were calculated from the skeleton bone rotations.

2.3.2 Gestures creation

Several eye, arms and hands movements were blended and combined with speech in order to make the avatar look more natural when interacting with the visitor. These movements created several gestures, including crossing arms, thinking, shrugging, raising hands and pointing (see figure 4).

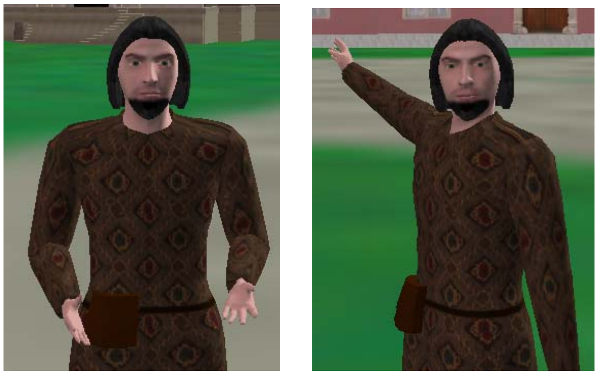


Figure 4: Raising hands and pointing gestures

2.3.3 Speech and lip synchronisation

Speech and lip movements were generated dynamically as the responses from the multilingual interactive avatar were produced in real time. Therefore, the lip movements of the avatar were created to look as realistic as possible. Unrealistic lip motion is distracting whilst realistic motions add to the immersive experience and aid those users who are hearing impaired but can lip read. The effectiveness of lip-reading and sensitivity to parameter settings remain to be evaluated.

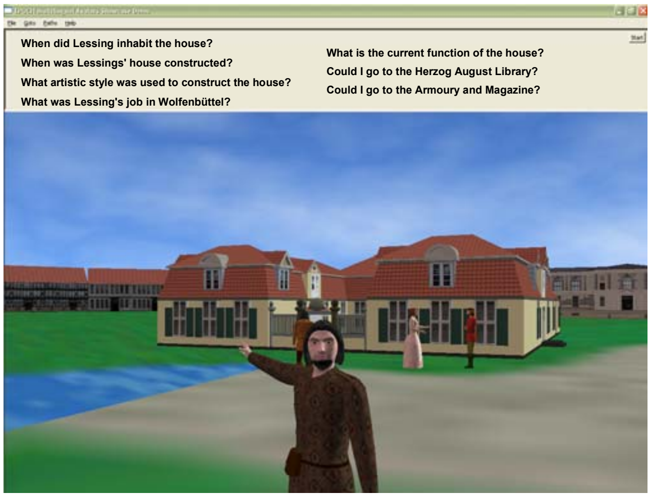
The speech was produced, by passing a script to the TTS system Vocalizer [10] from which a speech wave file was created. The script was also passed to a parser which produced a wordlist suitable for the recogniser. The recogniser is a simple phone based recogniser that quickly aligns the text and speech to produce a phone list with corresponding timings. This list was passed to the morph engine, which produced a set of facial morphs for each frame of animation.

2.4 Multilingual and multimodal language generation

The virtual environment was connected to a Multilingual and Multimodal Natural Language Generation (MMNLG) module based on technologies developed at Brighton [11]. This module was linked to a knowledge base of information with the user profile, such as age, language and the sites being visited. This knowledge was used to formulate in appropriate natural language, as well as synchronized movement and gestures, the avatar response to the visitor's questions.

The interaction between the visitor and the virtual environment follows the next sequence:

1. On arrival to a new location in the town, the system presents a list of questions that the visitor may wish to ask to receive more information about the current location, along with a list of questions as to which locations the visitor could go next (see figure 5).
2. The visitor selects a question using a typical ‘point&click’ operation and the virtual guide responds to the question in natural language giving more information about the location. In case the visitor has selected to move to another location, the guide moves to another location continuing the tour in the town.
3. When walking to another location, the virtual guide could stop several times to provide additional information related to the town and its famous inhabitants.
4. At arrival to the next location, a new set of questions is presented to the visitor and the interaction continues.



When did Lessing inhabit the house?
When was Lessings' house constructed?
What artistic style was used to construct the house?
What was Lessing's job in Wolfenbüttel?
What is the current function of the house?
Could I go to the Herzog August Library?
Could I go to the Armoury and Magazine?

Figure 5: Example of visitor interaction with the virtual environment

The architecture of the MMNLG module is shown in figure 6. The Dialogue Manager receives an input script with the user request, for example a request for an answer. Thereafter, the Content Selection module uses this request and the user’s profile, which is accessed from the user model, in order to select:

- An adequate answer to the question.
- A set of questions related to the location.
- The locations where the user has the option to go from the current location.

This information is sent to the Multilingual Surface Realisation module, which produces a scripted output with aligned gestures and movements. The input and output scripts between the MMNLG module and the VR environment and between the modules of the MMNLG were done using Rich Representation Language (RRL) language [12]. This is a rich formal language in XML suitable for representing the behaviour of conversational agents, which was developed within the NECA project [11].

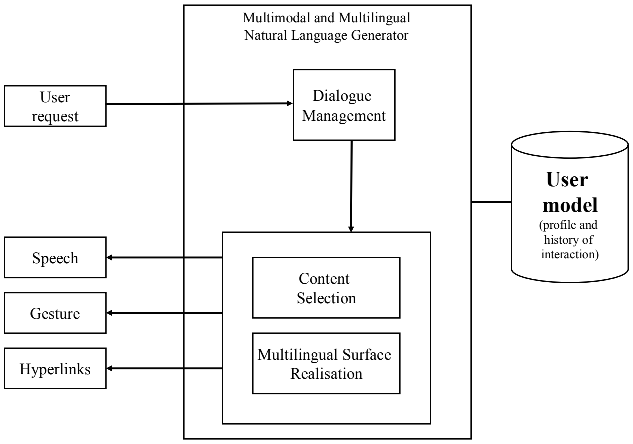

Multimodal and Multilingual Natural Language Generator
User request
Dialogue Management
User model
(profile and history of interaction)
Speech
Gesture
Hyperlinks
Content Selection
Multilingual Surface Realisation

Figure 6: Architecture of Multimodal and Multimodal Natural Language Generator

3. CONCLUSIONS

Building enjoyable and interactive virtual environments is essential for the understanding and appreciation of different European cultures. Moreover, with the advent of faster and cheaper hardware is increasingly more feasible to develop and take advantage of this type of environments. The modelling and rendering of Wolfenbüttel and its interactive avatars were designed for mid- to high-end consumer PCs; with support for multi-channel big screen theatres.

Nevertheless, there is still a need for further development of modelling and rendering techniques and other methods for achieving more natural and enjoyable interactions. This requires research in many areas, such as interactive technologies, advanced computers graphics, virtual humans, natural language analysis and generation, psychology and perception as well as usability studies.

4. FUTURE WORK

As a continuation of this work, a Novel Interfaces Laboratory at the University of Brighton is being built for conducting further research in building and using this type of interactive experience. This will involve not only research on the definition and real time display of complex virtual environments, but also usability evaluation of these interactive experiences. This will be achieved by experiments recording the use of interfaces and navigation systems using eye-tracking and other recording systems as well as by analysing effectiveness of techniques based on actual users and usage.

On the language side, we are pursuing vigorously the avenue of concept-to-speech generation, for the creation of affective and stylized output, using a freely available multilingual speech synthesizer [13]. This will allow the production of phones with their timing, avoiding the need for the error-prone and possibly slow speech recognizer in the generation pipeline. A task-oriented spoken German corpus and a domain-oriented spoken English corpus are currently being analyzed.

5. ACKNOWLEDGMENTS

This work has been conducted as part of the EPOCH network of excellence (IST-2002-507382) within the IST (Information Society Technologies) section of the Sixth Framework Programme of the European Commission. Our thanks go to the others involved in the realisation of this work: Phil Flack and Vince Jennings of the University of East Anglia and the team at the Technical University of Braunschweig, led by Prof Dieter Fellner and Sven Havemann.

6. REFERENCES

[1] Wojciechowski, R., Walczak, K., White, M., and Cellary, W. (2004) 'Building Virtual and Augmented Reality Museum Exhibitions', in Proceedings of 9th International Conference on 3D Web Technology, Monterey, USA. ACM SIGGRAPH. Pp. 135-144.

[2] Flack, P. and Willmott, J. and Browne, S. and Arnold, D.B. and Day, A.M., "Scene Assembly for Large Scale Urban Reconstructions", In VAST2001 Proceedings, Virtual Reality, Archaeology and Cultural Heritage, 2001

[3] Geser Guntram and Pereira John (2004) The Future Digital Heritage Space: An Expedition Report, Thematic Issue 7, DigiCult

[4] Androutsopoulos Ion, Kokkinaki Vassiliki, Dimitromanolaki Aggeliki, Calder Jo, Oberlander Jon and Not Elena (2001) "Generating Multilingual Personalized Descriptions of Museum Exhibits - The M-PIRO Project" In Proceedings of the 29th Conference on Computer Applications and Quantitative Methods in Archaeology, Sweden, 2001

[5] Gustafson, J., Bell, L., Boye, J., Lindström, A. and Wiren, M.: The NICE Fairy-tale Game System. Proceedings of SIGdial 04, Boston, 30 April-1 May 2004.

[6] S. Havemann, D.W. Fellner (2004) Generative Design of Gothic Window Tracery, In VAST2004 Proceedings, Virtual Reality, Archaeology and Cultural Heritage, 2004

[7] Willmott, J. and Wright, L. and Arnold, D.B. and Day, A.M., "Rendering of Large and Complex Urban Environments for Real Time Heritage Reconstruction", In VAST2001 Proceedings, Virtual Reality, Archaeology and Cultural Heritage, 2001

[8] Ekman, P., Friesen, W. V., Facial Action Coding System. http://www2.cs.cmu.edu/afs/cs/project/face/www/facs.htm

[9] Theobald, B., Matthews I., Bangham, J.A. and Cawley, G., 2.5D (2003) “Visual speech synthesis using appearance models”, Proceedings of the British Machine Vision Conference (BMVC), pp 43-52, Norwich, UK, 2003.

[10] Nuance (2005), http://www.nuance.com/

[11] Piwek, P. (2003). A Flexible Pragmatics-driven Language Generator for Animated Agents.Proc. of EACL03, Budapest, pp. 151-154

[12] Piwek Paul, Grice Martine, Krenn Brigitte, Baumann Stefan, Schröder Marc, Pirker Hannes (2002). RRL: A Rich Representation Language for the Description of Agent Behaviour in NECA. Proc. of the AAMAS workshop on Embodied conversational agents, Italy.

[13] MBROLA http://tcts.fpms.ac.be/synthesis/mbrola/