IST-2002- 507382

EPOCH

Excellence in Processing Open
Cultural Heritage

Network of Excellence

Information Society Technologies

D.2.4.4 (Final): Showcase 4 "Multilingual Avatars”

Due date of deliverable: 29 April 2005
Actual submission date: 28 April 2005

Start date of project: 15/03/2004
Duration: 4 Years

Ename Center

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

EPOCH Showcase 2.4.4
Multi-lingual Avatars

University of East Anglia
Technical University of Braunschweig
University of Brighton

1) Introduction

This showcase has been designed to illustrate the potential of combining low cost real time systems with technologies for modelling and rendering multi-lingual interactive avatars and populated urban scenes. It is clear that the historically common approach for creating 3D content is not ideal. This approach has been to create highly detailed, textured models with standard, all-purpose 3D modelling toolkits such as 3DStudio Max (Discreet) or Maya (Alias/Wavefront). These models are then exported to an all-purpose rendering engine to interactively explore the virtually reconstructed city.
The major problems with this approach are that because the modelling tools are general purpose the system is less able to exploit knowledge of the application domain to simplify the user interactions. In addition because models are created without exploiting domain knowledge they are more difficult to optimise for real time rendering of complex scenes. To put it another way – the person doing the modelling is very likely to create models which are inefficient to render and the system is less able to optimise these models if it has to assume that the user was creating completely general models.
More specialized modelling tools, which exploit knowledge of the types of object being modelling by working in the application domain, can be used to create appealing virtual reconstructions more quickly. At the same time, the structural information from the modeller gives valuable hints to the renderer to efficiently optimise interactive display through the use of level-of-detail and culling techniques. Thus a modeller that knows the operator is creating houses can use this information both to simplify the user interaction and guide the operator, and to create models that build in optimisations when attempting real-time rendering, exploiting the characteristics of the shapes.
In addition to good graphical rendering being essential to keep and attract the attention of the user, multimodal (and multilingual) systems need to process efficiently and integrate different modalities in such a way that the interaction appears natural and as human-like as possible. To be successful, interactive systems such as the one presented in this showcase shall include and integrate gesture representation and natural language understanding (ultimately spoken), as well as the generation of interesting route paths in the environment, lively and captivating gestures synchronized with meaningful (and possibly witty) natural language outputs. In other words, multimodality is concerned with the fusion of the input as well as with the fission of the output. In this showcase, priority was given to building a mock-up that will allow users with different background to be part of the full interactive loop (navigation-request-processing-response-navigation) of interaction with a virtual guide. To this end, multimodal input is limited to a selection in a menu of hyperlinks, while multimodal output combines gestures and speech generated on-the-fly in English and French. Taking into account cultural (language, age) as well as contextual (location on the site) information about the user during the interaction provides a first impression of what natural interactive systems can achieve for navigation through cultural heritage sites.

2) CHARISMATIC to EPOCH OpenSG conversion

The CHARISMATIC project created a toolkit to support the construction of large urban scenes with specially designed buildings, avatars, trees, and generic 3D objects. For this EPOCH showcase, two significant developments have been undertaken. Firstly, the scene construction technology has been extended. Secondly, the software has been converted to use OpenSG – an open source scene-graph framework – instead of using the tailor-made Charismatic scene toolkit, which had the advantage of efficiency but which was lacking in some areas.

The showcase demonstrator, which models the medieval German town of Wolfenbüttel, features the following objects:

- Skydome – taken directly from Charismatic;
- Terrain – direct from Charismatic, except it is now a static object, rather than a dynamic level-of-detail one, to take account of speed issues relating to OpenSG's bias against changeable geometry objects;
- Buildings – imported as simple geometry; Charismatic used OpenGL's Stencil Buffer for the creation of window openings, but OpenSG does not support this;
- Avatars – multiple-level-of-detail, walking, talking, interacting people.
- Crowd Avatars – avatars with low rendering overhead, to populate the background of the scene.

Scene Complexity Issues

Due to the large scale of the scene being created, methods of reducing the number of geometric features to be rendered in any given animation frame are needed in order to maintain interactive frame rates (25fps and above). The key technique used for the showcase demonstrator is that of Occluder Culling.

Occluder Culling

In urban environments there are many buildings positioned such that at ground level, every building in the scene potentially occludes other buildings and avatars in the scene. We take advantage of this fact by testing buildings and avatars against the occluder volumes of foreground buildings before they are rendered, with the aim of sending as small a number of objects as possible to the graphics card.

Occluder Lines are calculated for each building when it is loaded with the scene. For each building up to six lines are calculated – three in the xy-plane perpendicular to the z-axis and three in the yz-plane for the x-axis ('front' and 'side'). The algorithm to calculate these lines works as follows:

- Render the building in an orthographic projection with a black background and white for the building;
- Save the results of this rendering step in a buffer;
- From the ground level of the building (which may or may not be the bottom of the buffer, depending on the building), calculate the widest unbroken span of white pixels;
- Recursively move up the rendered image, one line at a time, finding at each stage the widest unbroken span whose end points do not extend beyond those of the span established so far;
- If the width does not reduce, replace the previous span with the new one;
- Each time the width reduces, store the span and start a new one.

Upon reaching the top of the building, find the best three spans with the following properties:

1) Biased to width (score = width*width * height, biggest is best);
2) Biased to height (score = width * height*height, biggest is best);
3) Biased to squareness (score = abs(1.0-(width/height)), smallest is best).

Some of the calculated occluders will be too short or narrow in relation to the size of the building for practical use, regardless of their score, so these are discarded.

Occluder Volumes

Each Occluder Line is associated with a tile in the scene – a typical scene consisting of 32x32 tiles. On every frame, the occluder lines in the tiles surrounding the camera are created as Occluder Volumes. Only those occluder lines within the current view frustum are used, and only those within the 25 tiles surrounding the camera. Occluder Lines too near the camera are ignored, as these cause problems when the camera is too close to a building.

The volume is created as shown in Figure 1.

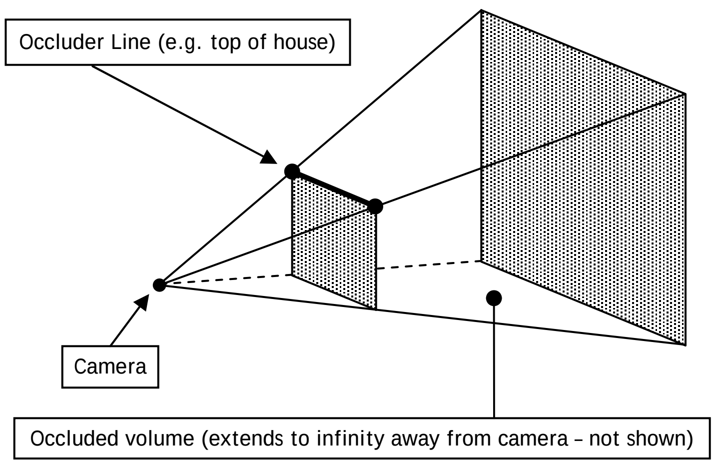


Figure 1: Occluder Volume

Each volume consists of four planes – top, left, right, and near. There is no bottom to the volume, as we assume that all buildings are sitting directly on the terrain. A near plane is not strictly necessary but it has been added to avoid the danger of close objects being incorrectly classified as occluded. An object qualifies as occluded only when its bounding volume is contained within all four planes.

Switch Nodes

Simple switch nodes are inserted into the scenegraph structure as a parent to each object that is potentially occludable – that is, above buildings and avatars. Before OpenSG renders the scene each node is tested to see if its children are occluded or not. If they are, the switch node is set to stop OpenSG traversing the children, thereby suppressing rendering of the given building or avatar).

Missing Techniques

Openings are stencilled holes in solid polygons which were used in the Charismatic toolkit to create windows and other openings in buildings. These are not used in the present system, as OpenSG does not natively support the stencil buffer. It was decided early on to not include this feature and to rely instead on OpenSG's fast polygon rendering.

People vs. the World

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 has been possible to ignore the problem of generating realistic animation of people walking on sloping or uneven terrain.

Another problem is that of collision detection between the moving people and buildings. Due to the large number of people in the crowds we have deliberately tried to avoid implementing full collision detection for every person as it would be too computationally expensive. Instead, a grid-based method is used, where each metre-square of the world is either accessible to the people or not. This works well for open spaces, but in narrow spaces, such as some of the streets in the Wolfenbüttel scene, this prevents people navigating freely. To stop them becoming stuck in these narrow passages we have manually closed them off, which could have been avoided with a full collision detection system.

3) Avatars

Work on this project involves porting avatars developed at the UEA into OpenSG which is targeted at large virtual environments for multi-screen rendering using clusters of PC's. The Wolfenbüttel scene is being modelled in the Charismatic virtual reality toolkit, which is also being ported to OpenSG, and populated by avatars which have multiple levels of detail and make use of image based techniques (see Figure 2). This allows representation of crowd scenes in the model, also developed at the UEA, as well as providing interactive avatars that can act as guides, complete with speech and face animation, that can respond to visitors' queries on specific buildings in the scene.

Figure 2: Wolfenbüttel scene

Morph target creation for a text to speech driven avatar.
The parameters derived from the speech recognition database are used to activate primitive morph targets on the avatar which are blended together to achieve the correct mouth shapes. The primitive shapes we are using are taken from the Facial Action Coding System (FACS) [1], which is a minimum set of mouth shapes based on muscle movement in the face (see figure 3). The shapes are specified by images and video, accompanied by descriptions.
At present, we have only implemented five parameters, one for jaw movement, one for tongue movement up and down, and three for the mouth. This set will expand as we derive more information from the database, and early results indicate that the final number of parameters could be less than 10.

Figure 3: Example mouth shapes

The morph targets are 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 are 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 are calculated in real time. From work carried out at the UEA [2], 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 are calculated from the skeleton bone rotations.

4) Speech

The virtual heritage sites modelled will contain avatars that can interact with users. These avatars can either be in a background crowd performing or acting as a guide at the front of the display. They will be multi-lingual.

The creation of new speaking avatars will be quick and easy to develop. The creation of morphs is a time consuming process, therefore the number of morphs needed to generate realistic looking speech will be kept to a small as number as possible. Multiple avatars in the scene should all be able to speak therefore the system will be able to accept multiple sources of speech. Some of the avatars will be interactive and their responses depend upon knowledge about the user and previous requests. Therefore some speech and lip movement need to be generated dynamically so the system has to be very resource efficient. Finally the lip movements should look as realistic as possible. Unrealistic lip motion will be distracting whilst realistic motions add to the immersive experience and aid those users who are hearing impaired but can lip read.

Previous work

Speaking avatars, often called talking heads, have been used in many applications. The principle design objectives are usually determined by the final use of the avatar. The SYNFACE project, [3, 4] was developed to aid the hearing impaired communicate over the telephone. A computer is used to generate synthetic lip movement for the incoming speech signal. This is done by using a speech recogniser to recognise the sequence of incoming phones. Each phone has an associated morph which is applied to the face. Using both the sound and synthetic lip movement, the users understanding of the speech was improved despite a phone error rate of 45%. Our task is easier in that we know the list of phones our avatar is supposed to enunciate and we simply need to time align them with the speech signal.

Architecture for generating synthetic lip movement

The architecture used by our system is shown in figure 4. A script is given to either an actor or a Text-To-Speech system from which a speech wave file is created. The script is also passed to a parser that produces 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 is passed to the morph engine that produces a set of facial morphs for each frame of animation.

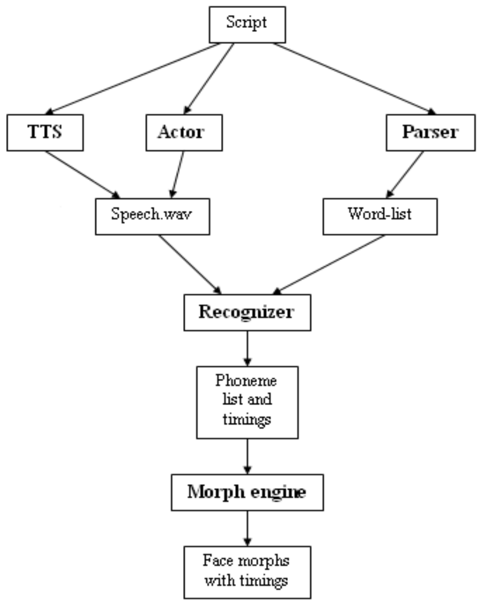


Figure 4: Architecture for lip movement generation

The morph engine takes a list of phone-states and their timings. These phone-states are used to generate a set of synthetic trajectories for the tongue tip, the jaw and the lips. These trajectories are based on data from the Mocha articulatory database [5]. From these trajectories the morphs are created. Figure 5 shows, (a) the reference face and six (b-g) the morphs used for visual speech synthesis. The amount of each morph required for each frame of animation is calculated based on the position of the articulators.

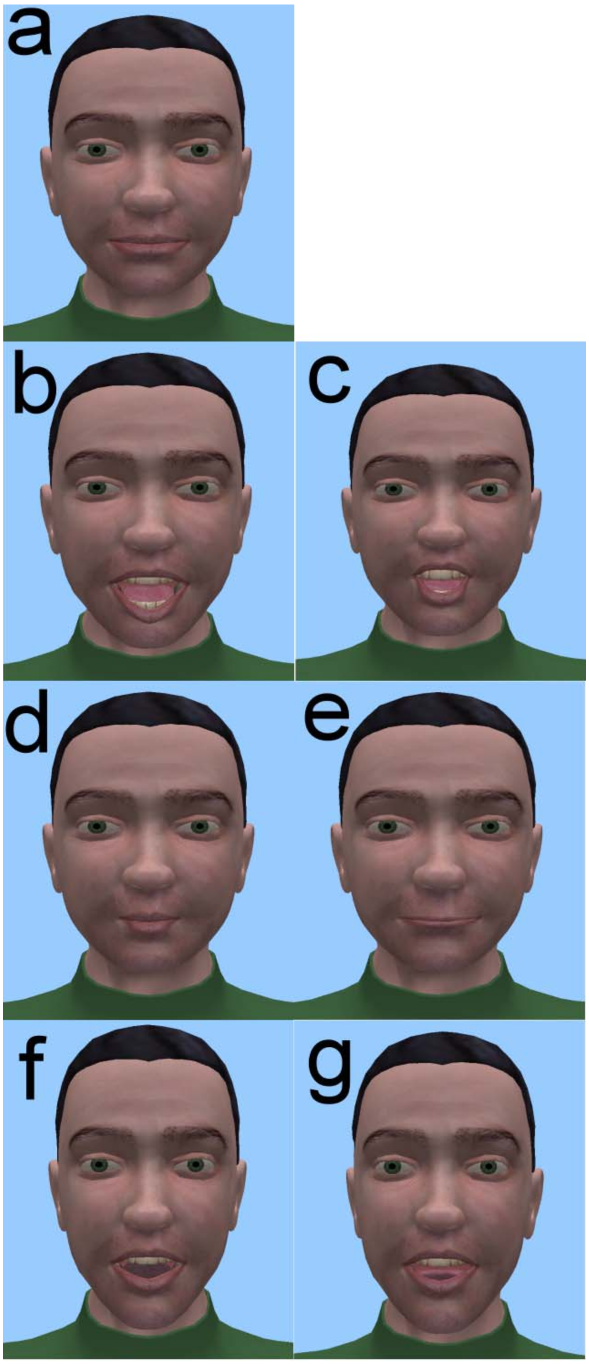


Figure 5: Reference face and six morphs

5) Multilingual and Multimodal Response Generation

This section presents a brief description of the work being carried out at the University of Brighton by ITRI with regards to the processing and generation of multimodal (speech and gestures) and multilingual interaction in a cultural heritage domain. The aim is to increase the accessibility of sites of cultural heritage by providing a virtual replica for visitors, with a guide who can speak more than one language and is sensitive to context (type of addressee, location, etc.).
To illustrate a typical interaction between the user and the Virtual Environment (VE), the Wolfenbüttel showcase involves mainly two types of situation and action:

On arrival to a new location, the VE displays a list of questions as to which locations the user can move, along with questions that the user may wish to ask about the current location. These questions are dynamically generated and adapted to the context of previously asked questions.
The user selects a question using a typical ‘point & click’ operation and the avatar moves to that location and/or gives information about the site.

A typical example of dialogue between the User (U) and the Guide (G) in the VE is shown in Figure 6. The example is in English, but the system also supports French. Questions are presented on a menu, where the selection is made by the user. Answers are automatically generated on the fly, in the user’s preferred language, and are presented as speech.

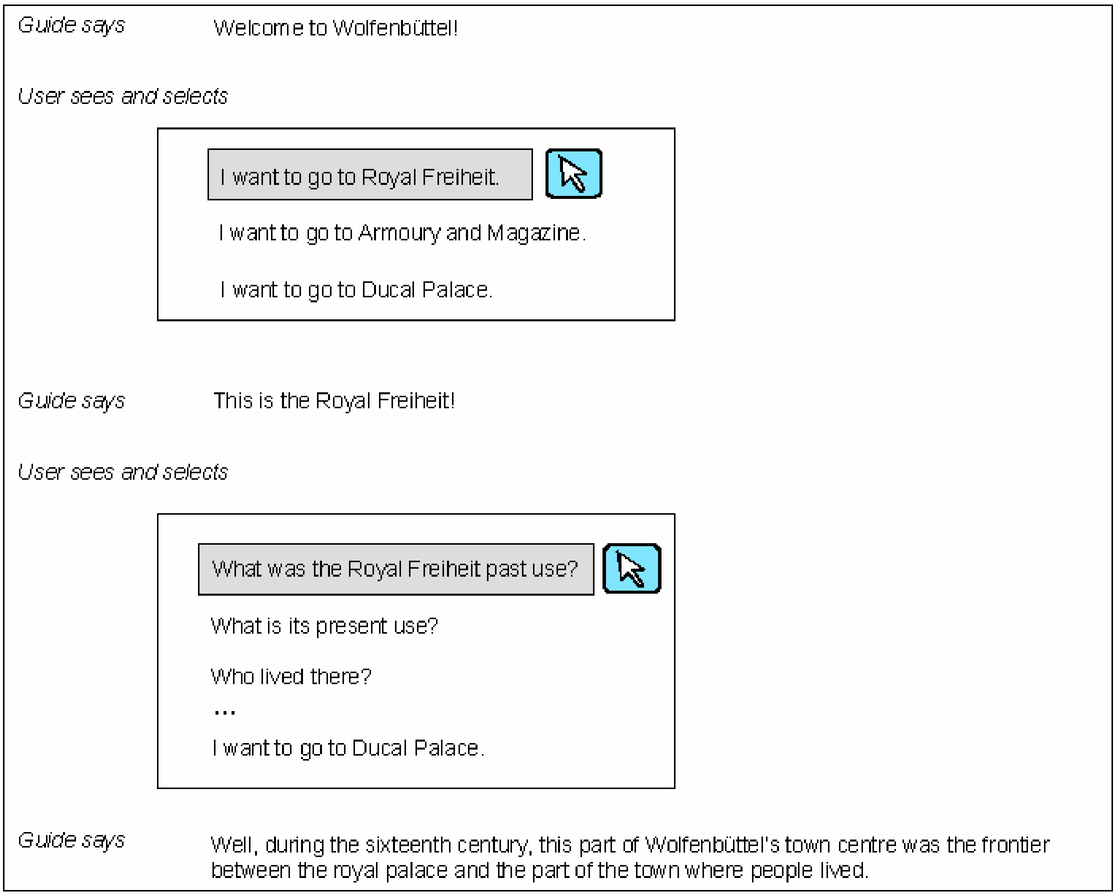


Figure 6: Interaction sample between user and guide

There are many interesting issues that this type of application addresses, and they are closely related to those involved with Human-Computer Interaction (HCI) systems, which are getting more and more popular as supporting technologies are getting better and become more widely available. Those issues involve tackling traditional problems in (multi-lingual) Computational Linguistics and Speech Processing (semantic interpretation of natural language input, neat generation of the output), but also more recent research areas such as Multimodality, Context-Sensitivity and the specific modelling of cultural heritage sites visitors.

To address those issues, ITRI has adopted the architecture shown in Figure 7. The showcase will feature a subset of the functionalities (the thick boxes) usually provided with HCI systems. The showcase will allow for a very simple mean of input interaction, namely hyperlink point and click operations, which means that Speech and Gesture inputs are not available at the moment, so that the Natural Language Understanding (NLU), Gesture Interpretation (GI) as well the Fusion modules are reduced to a single simple unit for input processing. Apart from linguistic knowledge, Dialogue Management (DM) uses four different types of knowledge. It must of course know about the site in which the tour is taking place, what kind of user it is addressing (age, language, origin), what is the history of the interaction up to that point and be aware of the scenery modifications, if any. The Multimodal and Multilingual Response Generator (MM-RG) produces speech output by using classical methods of Natural Language Generation (NLG) [Multilingual-Content Selection (M-CS) and Multilingual-Surface Realization (M-SR)]. Speech output is aligned with suitable gestures. The Multilingual-RG also produces a set of Hyperlinks (H-links) via a collection of predefined templates.

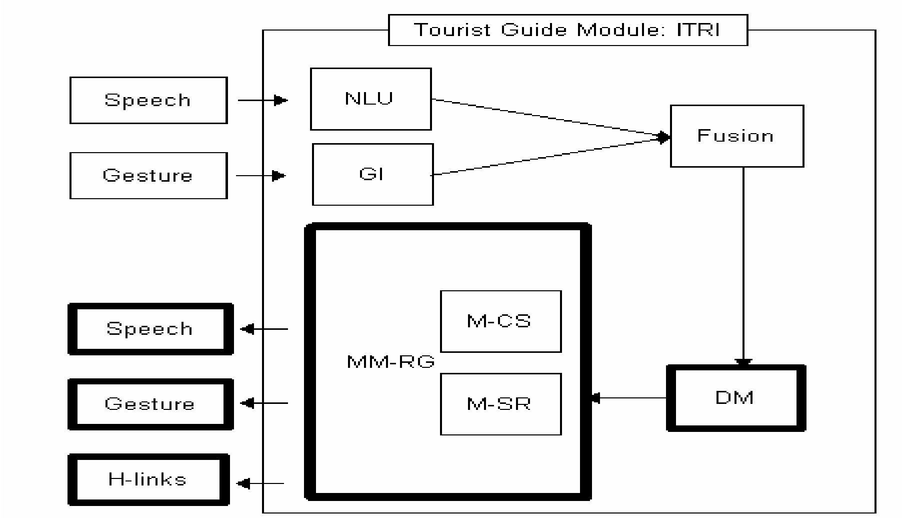
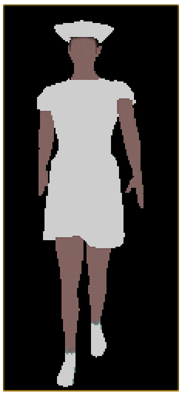


Figure 7: Architecture of Multimodal and Multimodal Data Processing at ITRI

6) Crowd Representation



8a) Example of a static texture used as a billboard

8b) A normal map (X = Red, Y=Green, Z=Blue)

8c) Example of a re-lit billboard

Without virtual humans in a virtual environment, the scene looks lifeless and the user is drawn to the details (or lack of) of the surrounding buildings. Adding a crowd brings visual interest to the scene, improves realism and draws attention away from other elements of the simulation. The goal is to add a crowd of virtual humans to a scene, incurring as minimal overhead as possible, while achieving acceptable visual results.

What are the particular problems associated with populating a scene with a crowd? As previously mentioned it is desirable that adding a crowd to a scene does not significantly impact upon the frame rate of the system. In addition to this, there is the requirement for the crowd to be visually varied. A crowd of identical humans would not be desirable, however even simple colour changes to clothes greatly enhance the visual variety.
The movement of the crowd is also significant to the simulation. Unrealistic actions, for example people walking through walls, are extremely distracting for a user. In addition to the problems of collision detection, the behaviour of the crowd in the general must be considered. The work, so far, has not focused on crowd dynamics; a simple behaviour system has been implemented at this stage.

Static billboard impostors, as proposed by Techia et al. [4] are used as the basis for this implementation. Static billboards have been chosen (Figure 8) due to the extremely lightweight rendering overhead per entity. The textures (colour and normal maps) are generated offline in a helper application and loaded into memory at the start of the main system.
At run-time the correct texture representation is chosen and via a fragment shader (implemented in OpenGL Shader Language (GLSL)) the texture is relit in the current lighting conditions using the information stored in the normal map. Due to the use of a fragment shader, the overhead of this relighting is proportional to the screen space occupied by each entity. This is advantageous for crowds, as on average each entity occupies a very small amount of screen space.

A height-map of the scene is maintained by the application. This serves two purposes, correct height positioning of the entities on the ground and also allowing virtual humans to detect (and thus avoid) obstacles. The crowd in this early stage is mainly for visual interest in the scene, the behaviour is a simple target point system for each entity.
There are several areas that deserve significant focus in future work. Firstly, we wish to examine the integration of the billboards with the geometry representations, with the aim of minimizing popping artefacts. This would require synchronisation of the animations, position and orientation of the models. Secondly the behavioural model needs significant work, including the integration of the work proposed in the paper by Ryder et al [5].

7) Future Research Relevant To This Showcase

As noted during both the Charismatic project and the creation of this showcase, one of the biggest problems in the visual appearance of the scenes is the lack of integration between the buildings and the terrain. For flat scenes this is not immediately apparent, but as soon as any elevated terrain meets a building the problem is visible. In order to overcome this, a method of integrating the vernacular buildings with the terrain is required. However, this method must conform to two criteria:

1. Preservation of the building geometry. Re-use of buildings without remodelling is a vital part of creating large urban environments.
2. Allowance for level-of-detail techniques, specifically on the terrain.

Future work on avatars will include linking mouth movement to activate morphs on other parts of the face, such as the cheeks and facial creases, determined by muscle actions, and the viability of controlling facial expression morphs, such as eyebrow movements, from speech intonation.

Gestures for avatars are pre-scripted animations that are combined with speech and facial gestures in real time in response to user input. This could be improved by using the Animgen software in the eSign project [8] which generates animations synthetically in real time, removing the need for specialist animation skills and providing a rich vocabulary of gestures and facial expressions. The avatars used in the showcase were initially developed for use in eSign. Currently for speech the synthetic articulatory trajectories generate realistic lip motions for most phone sequences. However this needs to be properly evaluated and improvements made where necessary.

A visual artefact of the current lip-synching is that the motion of the avatar’s lips appears jerky, despite the articulatory trajectories used to generate them being smooth. This is a consequence of the lips moving very quickly, the frame rate being relatively low at 25 fps and there being no motion blurring. This artefact could be relieved by increasing the frame rate or adding motion blur in future versions.

The mocha database contains information of the vertical and the forward/backward position of the lips but does not contain any information on lip spreading: this currently is inferred. A better approach would be to find a database containing this information and add it to the model. In addition all the data in the database is recorded on ‘Real speech’. ‘Real speech’ usually has more exaggerated intonation and prosody. This needs to be modelled to make the speech more realistic.

Probably the most advanced application that has been developed so far for virtual museum exhibits is M-PIRO [9] (Multilingual Personalised Information Object) and we believe that there are at least three specific areas in which ITRI could improve on similar systems. The current prototype allows only for input through hyperlinks, but allowing speech as an input modality, for which a trained parser, possibly stochastic, would developed specifically for museum visitors, would improve interactivity. Since multimodal generation is carried out via a system developed at ITRI called MNLG, which was developed during the NECA project [10], research effort can focus on the production of sophisticated output. Hence, the two remaining areas in which to pay much attention are Dialogue Management and Multilingual Natural Language Generation. Good Dialogue management makes it possible for users not only to get informative answers to their request, but to have entertaining conversations with the guide. Most of those issues have to do with how speech acts are selected from one turn to another, and recently, stochastic Dialogue Managers [11] have been developed for different domains, and ITRI may be considering such an avenue of research. Finally, NLG would greatly benefit from using concept-to-speech for generation, along with stylistic variations and randomization of the output.

References

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

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

[3] Granstrom, B., Karlsson, I., Spens, K.E: SYNFACE – a project presentation. 44 (2002), 93-99. Proceedings of Fonetik

[4] Karlsson, I., Faulkner A., Salvi, G. : SYNFACE – a talking face telephone. 2003, 1297-1300. Proceedings of Eurospeech.

[5] Wrench, A.: A multi-channel/multi-speaker articulatory database for continuous speech recognition research. PHONUS, 5 (2000), 1-13. 4

[6] Tecchia, F. and Chrysanthou, C. L. Y.: Visualizing crowds in real-time. Computer Graphics Forum 21, 4 (November 2002), 753–765
.
[7] Ryder, G., Flack, P. and Day, A. M. Adaptive Crowd Behaviour to Aid Real-Time Rendering of a Cultural Heritage Environment, Proceedings of VAST 2004.

[8] The Esign project
http://www.visicast.sys.uea.ac.uk/eSIGN/

[9] The M-PIRO project http://www.ltg.ed.ac.uk/mpiro/

[10] The NECA project http://www.ai.univie.ac.at/NECA/project/project.html

[11] Young, S. Talking to Machines (Statistically Speaking). Int. Conf. on Spoken Language Processing, Denver, Colorado (2002).

Multilingual Avatars

Cities and towns around the world have a huge wealth of historically significant and culturally important material in the form of buildings and events. We are creating a toolkit of applications specifically to enable this information to be recreated as virtual reality worlds for visitors to see for themselves how buildings once looked; to hear from virtual guides the history of significant buildings and events; to explore locations unavailable to them.

Modelling and rendering. A typical city scene might consist of a small (500 m^2) region with important buildings and a much larger surrounding region of generic housing. By providing tools to rapidly create these generic buildings and automatically position them

Wolfenbuttel

along roads, a large proportion of the scene can be constructed in a can be constructed in a relatively short time. A scene will also need important buildings (modelled using traditional modelling

packages) and parametric trees to add realism and movement. Once the background has been created a number of people can be added around the scene. These avatars can talk to the visitor and will be combined with a background crowd of city residents so that the model comes to life with an interactive information guide.

Recreation of a street in Wolfenbuttel

Speaking Avatars Speech is required to make interaction as natural and engaging as possible. The avatar must be able to respond to questions from a range of users with replies in a number of languages. Tools are required that allow new speaking avatars to be quickly developed. To this end, mouth shapes for all the speech sounds are created by mixing six base morphs and applying them to the reference face. Realistic visual speech can be generated using this limited set of morphs.

Technical Specifications The Wolfenbuttel showcase application is being developed using the open-source scenegraph OpenSG with custom add-ons for rendering speedup techniques including subdivision surface work from Braunschweig University.

The virtual environment needs to be related to a knowledge base of information about the site being visited, and this knowledge has to be formulated into the appropriate natural (text) language response, with synchronized movement and gestures as the visitor

asks questions of the guide.
Speech can be sourced from any Text to Speech Synthesizer or recorded by an actor. Phoneme timings are obtained by using the HTK recognizer. These phoneme timings are then used to create a series of morphs which drive the facial animation. Both modelling and rendering packages are designed for mid- to high-end consumer PCs; with support for multi-channel big screen theatres.

Partners This showcase is being realized by the following EPOCH partners. It will be demonstrated at the VAST 2004 conference.

- University of East Anglia, UK
- Technical University of Braunschweig, Germany
- University of Brighton, UK

Interested?

Are you interested in this showcase? Do you think that this approach can help you in creating effective Cultural Heritage presentation projects or can be integrated in new research projects? Please contact Dr. Andy Day (amd@cmp.uea.ac.uk) of University of East Anglia at +44 (0)1603 592604.

EPOCH is a Network of Excellence on Intelligent Cultural Heritage within the IST (Information Society Technologies) section of the Sixth Framework Programme of the European Commission. EPOCH showcases demonstrate innovative solutions and technological integration for target application areas in the Cultural Heritage domain. As they are created with real world content, they stimulate creative thinking about the use of the technologies in Cultural Heritage, and are used to validate new technological approaches with key stakeholders in the Cultural Heritage domain. For more details, visit the project web site:

www.epoch-net.org

EPOCH is funded by the European Commission under the Community's Sixth Framework Programme, contract no. 507382. However, this leaflet reflects only the authors' views and the European Community is not liable for any use that may be made of the information contained herein.