Image-based 3D modeling: modeling from reality

Luc Van Gool[1,2], Filip Defoort[1], Johannes Hug[2], Gregor Kalberer[2], Reinhard Koch[1],
Danny Martens[1], Marc Pollefeys[1], Marc Proesmans[1], Maarten Vergauwen[1], Alexey Zalesny[2]
[1] ESAT-PSI, Kath. Univ. Leuven, firstname.lastname@esat.kuleuven.ac.be
[2] IKT-BIWI, D-ELEK, ETH Zurich, lastname@vision.ee.ethz.ch

Abstract

Increasingly, realistic object, scene, and event modeling is based on image data rather than manual synthesis. The paper describes a system for visits to a virtual, 3D archeological site. One can navigate through this environment, with a virtual guide as companion. One can ask questions using natural, fluent speech. The guide will respond and will bring the visitor to the desired place. Simple answers are given as changes in the orientations of his head, by him raising his eyebrows or by head nodding. In the near future the head will speak.
The idea to model directly from images is applied in three subcomponents of this system. First, there are two systems for 3D modeling. One is a shape-from-video system, that turns multiple, uncalibrated images into realistic 3D models. This system was used to model the landscape and buildings of the site. The second projects a special pattern and was used to model smaller pieces, like statues and ornaments that often had intricate shapes. Secondly, the model of the scene is only as convincing as the texture by which it is covered. As it is impossible to keep images of the texture of a complete landscape, images of the natural surface were used to synthesise more of similar texture, starting from a very compact yet effective texture model. Thirdly, natural lip motions were learned from observed, 3D face dynamics. These will be used to animate the virtual guide in future versions of the system.

1 Introduction

We describe preliminary results for a virtual tour operator system. The demonstrator is centered around a visit to virtual Sagalassos, an ancient city in Turkey, that is being excavated by archaeologists of the University of Leuven. This demonstrator – coined EAMOS – integrates research on speech (Univ. Leuven) and vision (Univ. Leuven and ETH Zurich). The underlying, three-dimensional site model not only consists of the current landscape, ruins, and other finds, but it also contains CAD models of the original city. This helps to interpret the ruins in their original context.

One can navigate through this environment, with a virtual guide as companion. EAMOS allows a user to visually explore the scene with the assistance of this guide, which responds to spoken commands. The guiding agent presents itself as a hovering mask, and is able to communicate back to the user through head gestures and emotional expressions. The user is invited to inquire about the archeological site. The user is free to query for any interesting places to visit, or may ask for additional information about something visible in the scene. The guide takes the user on a tour, navigating from viewpoint to viewpoint in the scene. The visitors of virtual Sagalassos can formulate their queries through fluent, natural speech. The visual presence of the guide makes the interaction even more intuitive. As the mask reacts to the requests, a protocol is established that is quite similar to that of a normal person-to-person conversation. If wrongly understood, the user can soon pick up on the guide's mental state, as the facial mask frowns in anguish. Simple head nods confirm or negate questions, affirm or deny requests. In the near future the head will be animated to also let it speak, so that the guide can formulate more intelligent answers.

Fig. 1 shows some example views during such a virtual tour. The current line-up deployed for EAMOS consists of a single parallel processing computer (Onyx Infinite Reality), equipped with an audio interface. Different concurrently running software packages take care of rendering the scene, animating the guide, as well as processing and interpreting the speech.

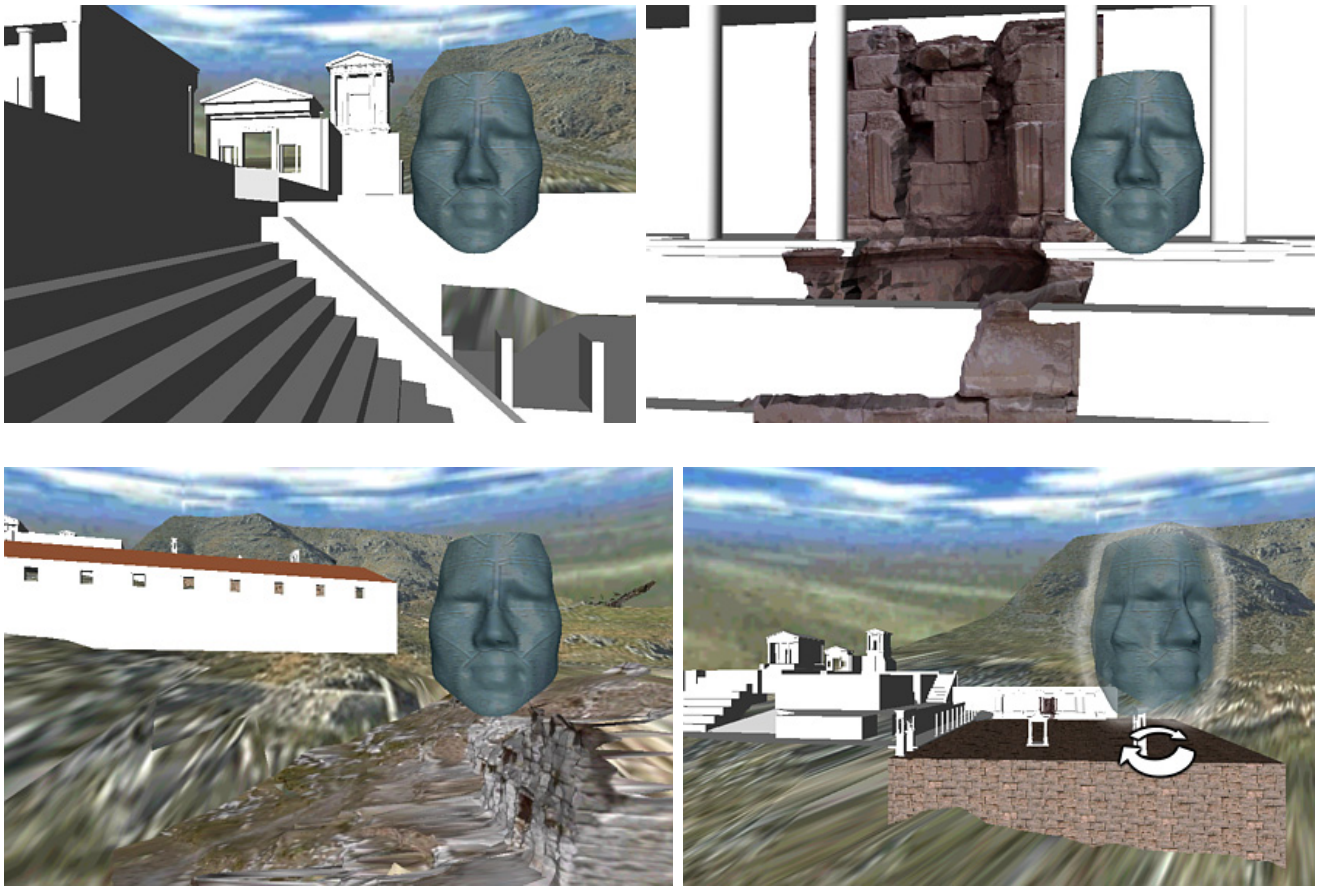


Figure 1. *Holiday pictures from a trip to virtual Sagalassos. The guide was very helpful, but not very talkative. The wheather was good: dry and room temperature throughout our stay. The figures show several of the components of which the site model is composed: a 3D landscape model, more detailed building models, CAD models that show the original shapes of the buildings and the context of the remaining ruins; the face mask of the virtual guide to whom questions can be asked in fluent speech and which reacts with emotional expressions.*

This contribution focuses on the vision, not the speech aspect. In particular it describes how the EAMOS demonstrator is based on three vision tasks, that each have been approached using image-based modeling as the primary paradigm:

1. 3D modeling of the landscape, ruins, and finds
2. modeling of the landscape texture
3. speech-oriented animation of the virtual guide's face

Currently, the team is integrating these techniques into a system that guides people around through virtual Sagalassos. A first version is ready, but each aspect needs improvement. Each of these aspects in now discussed in more detail, as well as our plans for the future.

2 Two image-based 3D acquisition systems

A first requirement for the EAMOS demonstrator is that visually convincing 3D models of the site be built. In the end, this will have to include a 3D model of the terrain (landscape), of the existing ruins, of the statuary (sculptures and ornaments), and of the different finds such as pottery. For now, initial models have been produced for the terrain, for some of the ruins, and for a few sculptures.

This section describes the two 3D acquisition systems that were used. They share the underlying idea of building systems that are easy to use and only require off-the-shelf hardware. This is important, as the archaeologists should be able to use the equipment *in situ* and without causing lengthy interruptions in the excavations. The systems should be brought to the finds and not *vice versa*. This is difficult if the acquisition equipment is either too expensive or too vulnerable. As it has to be carried around, the 3D acquisition systems should also be very light.

2.1 Shape-from-video

A first technique only requires a camera. It starts from multiple images, e.g. a video sequence. In contrast to traditional shape-from-motion or stereo approaches, the motion and intrinsic parameters of the camera are unknown. As a result, also existing footage can be used to reconstruct scenes that no longer exist. Much along the lines of work reported by Armstrong *et al.* [1], the method is based on the automatic tracking of image features over the different views. This is done in stages. First, a (Harris) corner detector is applied to yield a limited set of initial correspondences, which enable the process to put in place some geometric constraints (e.g. the epipolar lines as restricted search areas). These constraints support the correspondence search for a wider set of features and in the limit, for a dense, i.e. pixel-wise, field of disparities between the images [9].

The limited set of corner correspondences also yields the necessary data to perform a fully automated calibration of the camera and hence the camera projection matrices for its different, subsequent positions. Once these matrices are available, the 3D reconstruction of the observed scene can be produced. In general, to arrive at metric structure – i.e. to undo any remaining projective and affine skew from the 3D reconstruction – the camera intrinsic parameters like the focal length etc. have to remain fixed. But even if one has limited *a priori* knowledge about these parameters, like the pixel aspect ratio or the fact that rows and columns in the images are orthogonal, then also focal length can be allowed to change [7, 11, 12].

Fig. 2 gives an example of an historic building that has been reconstructed with this shape-from-video technique. It shows two of 6 images of an Indian temple, used for its 3D reconstruction. All images were taken from the same ground level as these two. Fig. 3 shows 2 views of the 3D reconstruction – a general overview and a detail – from viewpoints substantially different from those of the input images. The same method was applied to model the Sagalassos landscape. Several images were taken along the rim of a hill overlooking the excavation site. For several of the buildings (ruins) close range images were taken and also these were modeled. In all cases the intrinsic and extrinsic parameters of the camera were unknown.

As the method produces the list of intrinsic and extrinsic camera parameters one could also add virtual objects to the video sequences that were used as input. We have just started to explore such augmented reality work.

Our ongoing research in the shape-from-video area focuses on the following aspects:

1. to process longer image sequences fully automatically;
2. to integrate data from different sequences, e.g. for the exterior and interior of a building;
3. to solve the wide baseline correspondence problem, as to ensure that the system can automatically combine information from sequences taken from very different viewpoints;

Figure 2. *Two of 6 images of an Indian temple.*

4. to better combine the texture information contained in the different frames, e.g. to arrive at super-resolution;

5. to model different excavation strata in 3D and to integrate such information to build a detailed 3D, dynamic record of the excavations;

6. to extend the use of 3D acquisition technology to the support of virtual or real restoration and anastylosis, i.e. to use the 3D shapes of building blocks, sherds, and pieces in general to see how they can fit together. If the building or the artefact to which the pieces belong is of high scientific or artistic value a real restoration can then follow.

2.2 Active, one-shot 3D acquisition

The 'passive' technique outlined in the previous section cannot deal with untextured parts of a scene. This is a major problem with objects such as statues, the shape of which should be extracted with high precision, but which often do not have strongly textured surfaces. The same goes for the extraction of the shape of human faces, as is required for the animation of the guide's mask, as discussed later.

'Active' systems bypass the problem by projecting a pattern onto the scene. The 3D shape is extracted by analysing the displacements/deformations of the pattern when observed from a different direction (see [8] and Besl [2] for an overview). Typically such methods have relied on the projection of single points or lines and on scanning the scene to gradually build a 3D description point by point or line by line.

It is possible, however, to extract more complete 3D information from a single image by projecting a grid of lines. So far, such approaches had used additional constraints or information codes which force the grid to remain sparse [3, 13, 10]. With the technique we have developed and which has been refined and commercialised by Eyetronics [4] dense grids are projected, yielding high resolution 3D models. A single image yields the 3D shape of what is visible of the object to the camera and the projector. Fig. 4 shows the setup and a detail of an image from which 3D information can be extracted.

In order to also extract the surface texture, the lines of the grid are filtered out. Obviously, an alternative for static objects is to take another image without the grid. Yet, this is not an easy option if texture is to be obtained for dynamic scenes. The elimination of the grid is based on non-linear diffusion techniques and, of course, the precise knowledge of where the grid lines are in the image, but this is known from the shape extraction step.

Fig. 5 shows the 3D reconstruction of a Dionysos statue, found at the archaeological site of Sagalassos. It would be difficult to put such several tons heavy statue into the working volume of a laser scanner and it is not sure that the latter would survive...

Currently studied extensions to this technology include:

1. building a more compact, hand-held setup that is easy for use *in situ*;

2. the automatic crude registration of partial 3D patches, after which a traditional technique like ICP or mutual information is used to perform the fine-registration. Now crude registration still has to be done manually;

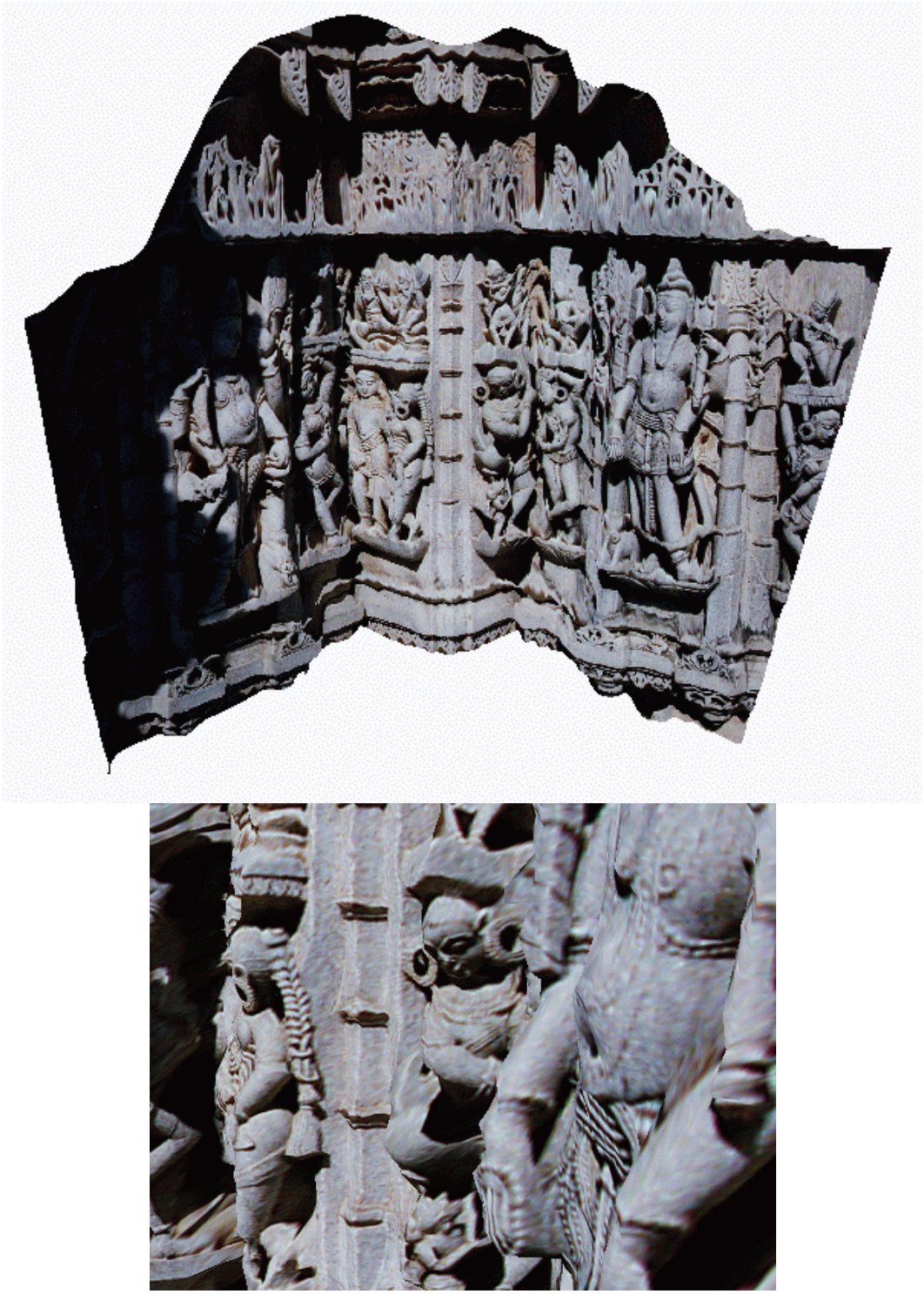
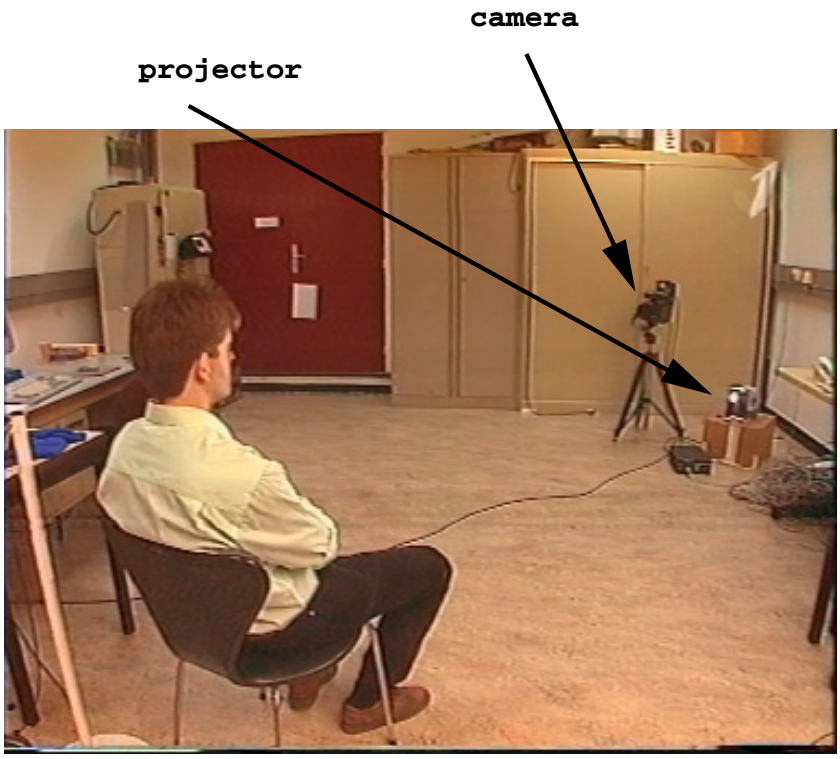


Figure 3. *Two views of the reconstruction obtained for the Indian temple.*



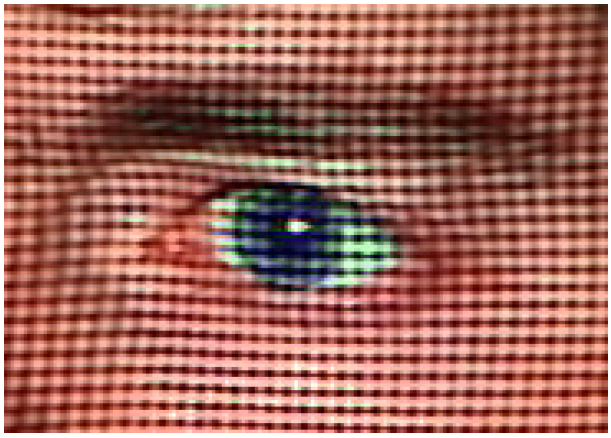


Figure 4. *Top: The active system only consists of a normal slide projector and camera, and a computer. The camera takes an image from a direction that is slightly different from the direction of projection. Bottom: A regular square pattern is projected on the scene, as seen in this detailed view. In this case, the grid covers the complete face. 3D coordinates are calculated for all the line intersections, resulting in the simultaneous measurement for thousands of points.*

Figure 5. *Two views of the reconstructed Dionysos statue.*

Figure 6. *Image showing terrain texture at the Sagalassos site.*

3. to specialise the setup also for pottery sherds, which are very important in archaeology for dating the stratigraphic layers uncovered during the excavations.

3 Image-based texture synthesis

Only a rather rough model of the landscape has been built. In fact, the resolution of this model is much coarser than that of some of the ruins. As one moves from building to building and crosses the bare landscape in between, there is a noticeable and disturbing difference in visual detail between the textures. On the other hand, precise modeling of the landscape texture would cost an enormous amount of time and memory space. Also, such precise modeling is not really required. It would for most practical means suffice to cover the landscape with a texture that looks detailed and realistic, but that does not necessarily correspond to the real texture on that particular part of the site. Thus, as a compromise we model the terrain texture on the basis of a few, selected example images. Such example image is shown in fig. 6. The resulting model is very compact and can be used to generate arbitrarily large patches of texture that look very similar to the exemplar texture. Emphasis so far has been on the quality of the results rather than the efficiency of the texture synthesis.

The approach builds on the cooccurrence principle: nearly all possible pairwise interactions in the example texture image are analysed. The fact that only pairwise interactions are analysed is in line with Julesz's observation that mainly first and second-order statistics govern our perception of textures. Yet, it is well-known that third and higher order statistics cannot be neglected just like that, mainly because of figural patterns that are not preserved. Here we are dealing with natural textures

and this issue is less crucial. Nevertheless, this restriction is dictated rather by the computational complexity and not by the underlying principles.

Textures are synthesised as to mimick the pairwise statistics of the example texture. Just using all pairwise interactions in the model is not a viable approach and a good selection needs to be made [6]. We have opted for an approach that makes a selection as to keep this set minimal but on the other hand bring the statistics of the synthesised textures very close to that of the example textures [14]. Parameter selection follows an iterative approach, where pairwise interactions are added one by one, the synthetic texture is each time updated accordingly, and the statistical difference between example and synthetised texture is analysed to decide on which further addition to make. The set of pairwise interactions selected for the model (from which textures are synthesised) is called the neigbhourhood system.

A sketch of the algorithm is as follows:

step 1: Collect the complete 2nd-order statistics for the example texture, i.e. the statistics of all pairwise interactions. (After this step the example texture is no longer needed) As a matter of fact, the current implementation doesn't start from all pairwise interactions, as it focuses on interactions between positions within a maximal distance.

step 2: Generate an image filled with independent noise and with values uniformly distributed in the range of the example texture. This noise image serves as the initial synthesised texture, to be refined in subsequent steps.

step 3: Collect the full pairwise statistics for the current synthesised image.

step 4: For each type of pairwise interaction, compare the statistics of the example texture and the synthesised texture and calculate their 'distance'. For the statistics the intensity difference distribution (normalised histograms) were used and the distance was simply Euclidean. In fact, the intensity distribution of the images was added also, where 'singletons' played the role of an additional interaction. The current implementation uses image quantization with 32 gray levels.

step 5: Select the interaction type with the maximal distance (cf. step 4). If this distance is less than some threshold go to step 8 – the end of the algorithm. Otherwise add the interaction type to the current (initially empty) neighborhood system and all its statistical characteristics to the current (initially empty) texture parameter set.

step 6: Synthesize a new texture using the updated neighourhood system and texture parameter set

step 7: Go to step 3.

step 8: End of the algorithm.

For texture synthesis the images are treated as a realization from the family of Markov random fields with the neighborhood system corresponding to the selected interaction types. The convergence of the corresponding relaxation procedure to a single stationary point has been proven [14].

After the 8-step analysis algorithm we have the final neighborhood system of the texture and its parameter set. This model is very small compared to the complete 2nd-order statistics extracted in step 1. Typically only 10 to 40 pairwise interactions are included and the model amounts from a few hundreds to a few thousands bytes. Nevertheless, it yields small statistical differences between the example and synthesised textures.

This texture synthesis approach can handle quite broad classes of textures. Nevertheless, it has problems with capturing complex semantic orderings or texels of texels with specific shapes. The method has mainly been used for coloured textures, as is also required for the Sagalassos virtual site. In the case of colour images pairwise interactions are added that combine intensities of different bands. The shortest 4-neighborhood system and the vertical interband interactions were always preselected because experiments showed that they are important for the vast majority of the texture classes. Fig. 7 shows a synthesised textured for the example image in fig. 6. Fig. 8 shows part of the site with the original terrain model texture (left) and with synthesised texture mapped onto the landscape (right).

Ongoing work is aimed at extending the results in order to

1. include the 3D nature of texture: the idea is to model textures from images taken from different views, but without a complete extraction of the neighbourhood system for every view separately as this would take too much time;

2. compress the 3D texture models: this could be done by exploiting the relation between a texture's appearance for different viewing angles;

Figure 7. *A synthesised texture based on the example image of fig. 6.*

Figure 8. *View of the old bath house and surrounding landscape at Sagalassos. Left: view with the original landscape texture. As this is a view strongly zooms in onto this model, the texture is of insufficient quality. Right: the landscape texture has been replaced by synthetic texture.*

3. to achieve viewpoint consistency: if the goal is to move around in a scene, the texture at a certain location should change in a way that is consistent with the texture generated in the previous views, e.g. pattern mimicking rocks should not be shifted around. The hope is to achieve this by drving the probabilitiex for the generation of different colour patterns not only from the model for the required viewing angle, but also from transition probabilities given the previous view.

4 Face animation for speech

Currently the guide only listens and answers through facial expressions, but he doesn't talk. Work on face animation should change this. The plan is to learn realistic 3D representations of visemes from observed 3D lip motions captured with the active system (section 2.2). As 3D can be captured from a single image, one can also take a video of a moving or deforming object and get as many 3D reconstructions as there are frames. Fig. 9 shows the 3D reconstructions extracted from three frames of a talking head video, each seen from three different viewpoints. From video data taken with a normal camera

Figure 9. *3D reconstructions of a face for three frames of a video and shown from 3 different viewpoints.*

25 (or 30) reconstructions can be made for every second of motion. The quality suffices to carry out detailed investigations into 3D face dynamics. Each 3D snapshot consists of 3D data for thousands of points (the full grid containss 600×600 lines and for every intersection a 3D coordinate can be given out by the system, so camera resolution is the limiting factor here). The 3D reconstructions can be made at the temporal resolution of the video camera, but processing is done off-line. For the moment, the reconstruction of a single frame incl. texture takes about 2 minutes.

At the time of writing, 3D dynamics have been captured for a basic set of 16 visemes, following conclusions by Ezzat and Poggio [5]. In a first step, a topological lip mask was fitted to the different 3D mouth positions. This mask is illustrated in fig. 10.

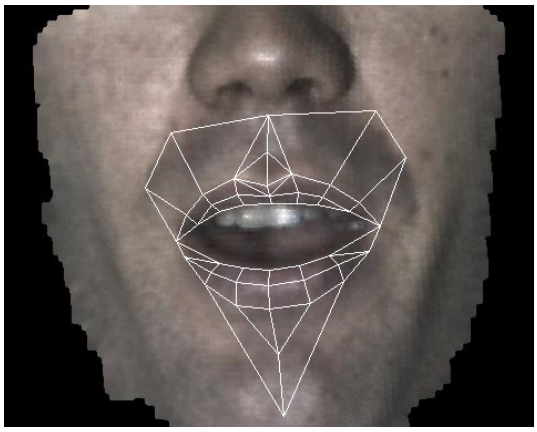


Figure 10. *A lip topology mask is used to support tracking.*

Statistics were extracted for the mask nodes positions. These were used to generate a robust lip tracker. Apart from the 3D positions, the tracker also uses colour information and 3D surface curvature. Fig. 11 shows the lip mask as it was automatically fitted to different mouth poses, in order to learn 3D lip dynamics for the 16 basic visemes.

The work on face animation has just started. Several issues are under investigation:

1. the set of basic visemes, including co-articulation effects, needs to be determined. Currently, there is not much agreement on this point in the literature;
2. these visemes have to be extracted with a high degree of automation from examples for different people, in order to draw the necessary statistics;
3. the step from analysis to synthesis/animation of the virtual guide has to be performed, based on speech input.

5 Conclusions and future work

In this paper we dicussed ongoing work on a system that guides visitors through a virtual archaeological site. The underlying 3D acquisition technology was concisely described. It plays a crucial role in making such large-scale projects possible, as it is easy to operate and yet yields realistically lookimg models. A similar philosophy of modeling from observations was used to synthesise textures similar to those found on site and to learn 3D mouth dynamics for a range of visemes.

Much work remains to be done also on the visualisation side. In future implementations, EAMOS will try to anticipate user requests through user modeling. Also will more 3D models be produced, of additional buildings (with the passive shape-from-video technique) and finds (with the active one-shot technique). Also will the CAD reconstructions be extended to represent different periods: from the 2nd century BC (Greek period, Sagalassos' heydays) up to the 6th century AD (Christian period, decline and shortly before its destruction by an earthquake, after which it was abandoned for good). In parallel, some of our colleagues are working on the compression of the 3D models and level-of-detail oriented compression.

Acknowledgements: The authors gratefully acknowledge support of ACTS project AC074 'VANGUARD', IUAP project 'Imechs' financed by the Belgian OSTC (Services of the Prime Minister, Belgian Federal Services for Scientific, Technical, and Cultural Affairs) and GOA project 'HVS' sponsered by the K.U.Leuven Research Council. Filip Defoort and Marc Pollefeys gratefully acknowledges support of a FWO grant (Flemish Fund for Scientific Research). Alexey Zalesny is supported as 'Akademischer Gast' by the ETH. The authors also gratefully acknowledge support by ETH through the 'Visemes' project. Finally, the authors also want to thank their colleagues from the speech group at PSI, K.U.Leuven for providing the natural speech understanding system.

References

[1] M. Armstrong, A. Zisserman, and P. Beardsley, Euclidean structure from uncalibrated images, Proc. British Machine Vision Conf., 1994

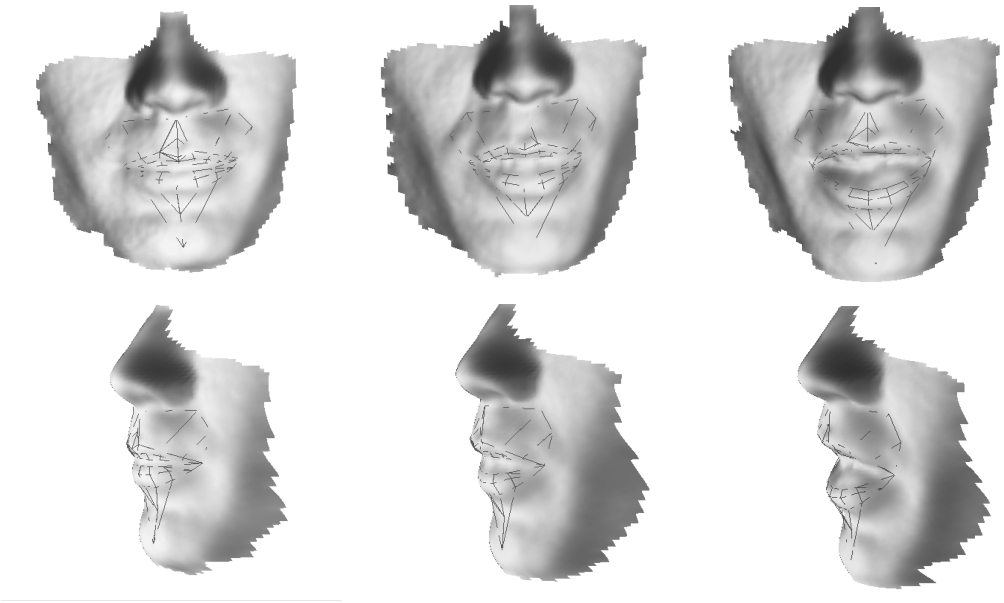


Figure 11. *The lip topology mask is fitted automatically to the 3D data, using 3D positions, surface colour, and surface curvature.*

[2] Besl, P., Active Optical Range Imaging Sensors, Machine Vision and Applications, Vol. 1, No. 2, pp.127-152, 1988

[3] K. Boyer and A. Kak, Color-encoded structured light for rapid active ranging, IEEE Trans. Pattern Anal. and Machine Intell., Vol. 9, No. 10, pp. 14-28, 1987

[4] http://www.eyetronics.com

[5] T. Ezzat and T. Poggio, Visual speech synthesis by morphing visemes, AI Memo No. 1658, MIT, May 1999

[6] A. Gagalowicz and S. D. Ma, Sequential Synthesis of Natural Textures. CVGIP, vol. 30, pp. 289-315, 1985

[7] A. Heyden and K. Aström, Euclidean reconstruction from image sequences with varying and unknown focal length and principal point, Proc. Conf. on Computer Vision and Pattern Recognition, pp. 438-443, 1997

[8] Jarvis, A persepctive on range finding techniques for computer vision, IEEE Trans. on PAMI, Vol. 5, No 2, pp.122-139, 1983

[9] R. Koch, M. Pollefeys, and L. Van Gool, Multi viewpoint stereo from uncalibrated video sequences, Proc. Eur. Conf. Computer Vision, Vol.I, pp. 55-71, 1998

[10] M. Maruyama, and S. Abe, Range Sensing by Projecting Multiple Slits with Random Cuts, IEEE PAMI 15(6), pp. 647-650, 1993.

[11] M. Pollefeys, R. Koch, and L. Van Gool, Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters, Int. Conf. on Computer Vision, pp. 90-95, Bombay, India, jan. 4-7, 1998

[12] M. Pollefeys, R. Koch, and L. Van Gool, Self-calibration and metric reconstruction inspite of varying and inknown intrinsic camera parameters, Int. Journal of Computer Vision, Vol.32, No.1, pp. 7-25, 1999

[13] P. Vuylsteke and A. Oosterlinck, Range Image Acquisition with a Single Binary-Encoded Light Pattern, IEEE PAMI 12(2), pp. 148-164, 1990.

[14] A. Zalesny, Analysis and Synthesis of Textures With Pairwise Signal Interactions. Techn. rep. KUL/ESAT/PSI/9902, Katholieke Universiteit Leuven, Belgium, 1999, 47p.