NATO workshop Stirling 1997
Face recognition : from Theory to Applications.

Getting Facial Features and Gestures in 3D

Marc Proesmans and Luc Van Gool

Kath.Univ.Leuven, Kard. Mercierlaan 94, 3001 Leuven, Belgium
Marc.Proesmans@esat.kuleuven.ac.be

Abstract

In the study of face perception and face animation, there is a growing interest in using 3-D data. The advent of dedicated laser scanners has proved instrumental in this regard. The availability of 3D face models makes it easier to change viewpoints and illumination conditions in perception experiments or to build animated likenesses in graphics. Here we propose to go one step further and also capture face dynamics (visemes, epressions) in 3-D. To that end, an active 3-D acquisition system is proposed, that yields 3-D, textured snapshots from a single image. By the fact that data are captured from a single shot, the face may move during the operation. Alternatively, images can be taken at video rate and for each frame a 3-D reconstruction (still textured if required) can be made. Such 3-D movies can then be used to analyse facial expressions in 3-D, through the tracking of points or features on the face. Preliminary experiments in that direction are presented. The reported work is part of an effort to model facial expressions without taking recourse to the modeling of the underlying physiology.

reseach.

1 Introduction

An ideal authentication system should allow quite some liberty in the head pose of the person to be checked. Similarly, it would be ideal if the person can be in motion during the control. Insisting on the person freezing before the system would make it rather intrusive or offensive, or at least a bit of a nuisance. It seems easier to reach such goals with 3-D measurements of the face than when only 2-D images are taken. From a 3-D description it is easier to extract viewpoint-invariant characteristics. In order to deal with moving faces, traditional active devices would pose difficulties, however, as they require scanning operations of several seconds. Moreover, only a few of such systems yield the surface texture, which is crucial for face recognition by humans. It stands to reason then to add texture extraction for the recognition of faces by computers.

Getting Facial Features and Gestures in 3D

Marc Proesmans and Luc Van Gool

Katholieke Universiteit Leuven,
Kard. Mercierlaan 94, 3001 Leuven, Belgium

Abstract In the study of face perception and face animation, there is a growing interest in using 3-D data. The advent of dedicated laser scanners has proved instrumental in this regard. The availability of 3D face models makes it easier to change viewpoints and illumination conditions in perception experiments or to build animated likenesses in graphics. Here we propose to go one step further and also capture face dynamics (visemes, epressions) in 3-D. To that end, an active 3-D acquisition system is proposed, that yields 3-D, textured snapshots from a single image. By the fact that data are captured from a single shot, the face may move during the operation. Alternatively, images can be taken at video rate and for each frame a 3-D reconstruction (still textured if required) can be made. Such 3-D movies can then be used to analyse facial expressions in 3-D, through the tracking of points or features on the face. Preliminary experiments in that direction are presented. The reported work is part of an effort to model facial expressions without taking recourse to the modeling of the underlying physiology.

1 Introduction

An ideal authentication system should allow quite some liberty in the head pose of the person to be checked. Similarly, it would be ideal if the person can be in motion during the control. Insisting on the person freezing before the system would make it rather intrusive or offensive, or at least a bit of a nuisance. It seems easier to reach such goals with 3-D measurements of the face than when only 2-D images are taken. From a 3-D description it is easier to extract viewpoint-invariant characteristics. In order to deal with moving faces, traditional active devices would pose difficulties, however, as they require scanning operations of several seconds. Moreover, only a few of such systems yield the surface texture, which is crucial for face recognition by humans. It stands to reason then to add texture extraction for the recognition of faces by computers.

In the case of face animation for virtual actors, avatars, videophone talking heads, and diverse applications of graphics, the importance of 3-D, textured face descriptions has also been recognised. Starting from such data, efforts have been made to build anatomical models, with detailed models for the skin, muscles and skull [13]. Animation is then guided by the physiological processes that govern the dynamics of this skin/muscle/skull complex. Sometimes video data are used to drive such model, based on the 2-D tracking of features. The question is whether a more phenomenological approach couldn't be more beneficial, due to its simplicity. At the end of the day, humans don't have to know about muscles pulling, pushing, and squeezing skin to 'know' all too well whether a facial expression looks natural or not. Thus, it seems interesting to model facial dynamics directly from observations. This requires measuring the 3-D changes that faces undergo during expressions and this with a sufficiently frequent, temporal sampling. Even if in the end intermediate views would be generated by the interpolation of extremal positions of the facial features, knowledge about intermediate stages may suggest the most appropriate interpolation schemes.

The paper proposes a 3-D acquisition method that fullfills several of these requirements. It extracts 3-D information ánd the corresponding surface texture from image data that can be captured at video-rate. This is the basis for its capacity to densely sample the changes that 3-D shapes undergo over time.

In section 2 the 3-D acquisition system is described. Section 3 illustrates the use of the system for the extraction of 3-D face shape and texture. Section 4 describes a first application, where the data are used to assist police forces in the identification of offenders. This work includes the derivation of a simple model for skin reflectance. Section 5 shows the use of the 3-D data for the extraction and tracking of facial features such as the lips, the mouth and the nose. This is useful e.g. to animate a virtual face from the facial expressions of an actor. The section illustrates this with a simple 'special effect' movie.

2 Three-dimensional face capture

Extracting 3-D information from scenes is a longstanding research issue in the computer vision community [9]. Two major strands have developed.

A first class are so-called 'passive' techniques, that work with normal, ambient light. The reconstruction is based on finding the projection of several points in different images taken from different viewpoints. If the images are taken simultaneously, the scene may contain moving parts and the motion can even be retrieved by processing subsequent camera frames [21]. Moreover, given the 3-D shape and the multiple views, one is in a good position to map texture onto the surface or to try and extract the surface reflectance

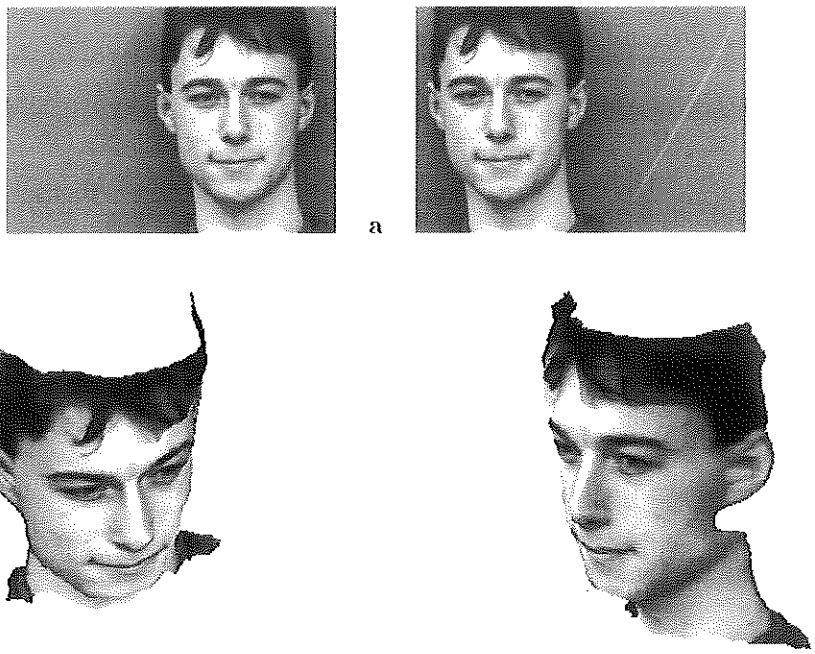


Figure 1: **a:** *Original stereo image pair of a face.* **b:** *Two views of the reconstructed surface.*

characteristics in detail.
Passive techniques also pose problems, however. First, the correspondence problem, i.e. the search for the same points in several views, is still a weak but essential step. Even with the latest algorithms the precision of passive, 3-D reconstructions compares unfavourably with the active techniques discussed next. Figure 1 shows the results for a face using a passive technique [18]. Although convincing to some extent, accuracy is wanting when it comes to person identification or the modelling of well-known persons.

A second class of 3-D acquisition systems are 'active' in the sense that they apply a special illumination to extract the 3-D information (for a review see e.g. [1, 9]). The illumination should reduce the problem of finding correspondences. The projected patterns are often observed with two or more cameras, but also from a single view can 3-D information be gathered, with the projector replacing one of the cameras in a stereo system. On the whole, active techniques yield higher precision. Simultaneously, the image processing operations are simpler. They obviously are also more intrusive, although the pattern(s) could be projected using near-infrared light, to which CCDs

are quite sensitive. The projection of the pattern can also make it more difficult to extract the underlying surface texture. As a consequence, active devices can yield misalignments between the shape and its texture, because they are captured neither simultaneously nor by the same sensor. Also – and this is particularly important for faces – active systems usually scan the object surface or use a series of subsequent projections. In such cases, the acquisition time easily gets too long to deal with object motion.

The system proposed here combines features of both approaches, making it quite appropriate for 3-D face capture. It uses a special illumination pattern to obtain good precision, but only a single image is used for the extraction of the 3-D shape ('one-shot' operation). From the same image alsso the surface texture is extracted, resulting in a perfect alignment between the two. This one-shot system is an improvement over an earlier version by our lab [23], where the spatial resolution was lower, the system more difficult to set up, and surface texture was not extracted. The improvement was possible because the illumination pattern was simplified. Fig. 2(a) shows the setup. Note that the projector and camera can be put quite close, leaving a small opening angle between the rays of viewing and projection. The advantage is that problems of occlusion are minimised and 3-D reconstruction can be performed until close to the occluding boundaries of the object's surface (as viewed from the camera). Fig. 2(b) shows a detail of the face, so that the pattern is clearly visible. It consists of a simple line grid.

¿From the perspective of active systems, an interesting novelty it that both 3-D shape and texture are extracted from a single image. As already mentioned, there is no alignment problem between the two because they are derived from the same image. Fig. 3(a) shows an image of a face, (b) the extracted 3-D shape, and (c) the result with the image texture mapped. Texture extraction is based on filtering out the grid lines from the original image or, if the person keeps very still, from a second image taken without the pattern. In all the examples shown in the paper, the former approach – filtering out the lines – was used.

The system is not actually a 'range acquisition' system. So far, 3D acquisition has been almost synonymous to range – i.e. distance – extraction. Getting shape via variations in distance might be a costly detour, however. It usually is the requirement to know absolute distance that complicates the necessary hardware and calibration. The system yields 3-D shape only up to scale. This can be easily fixed by giving a single, measured length, if there is a need...

It is also useful to note that this system is easy to calibrate. It suffices to show a scene with two dominant planes (like the corner of the room or a box for instance) and to specify the angle subtended between these planes. From there, the system autocalibrates. This makes it easy to change the setup of the camera and the projector or, generally, to transport the system toward the people or objects to be reconstructed.

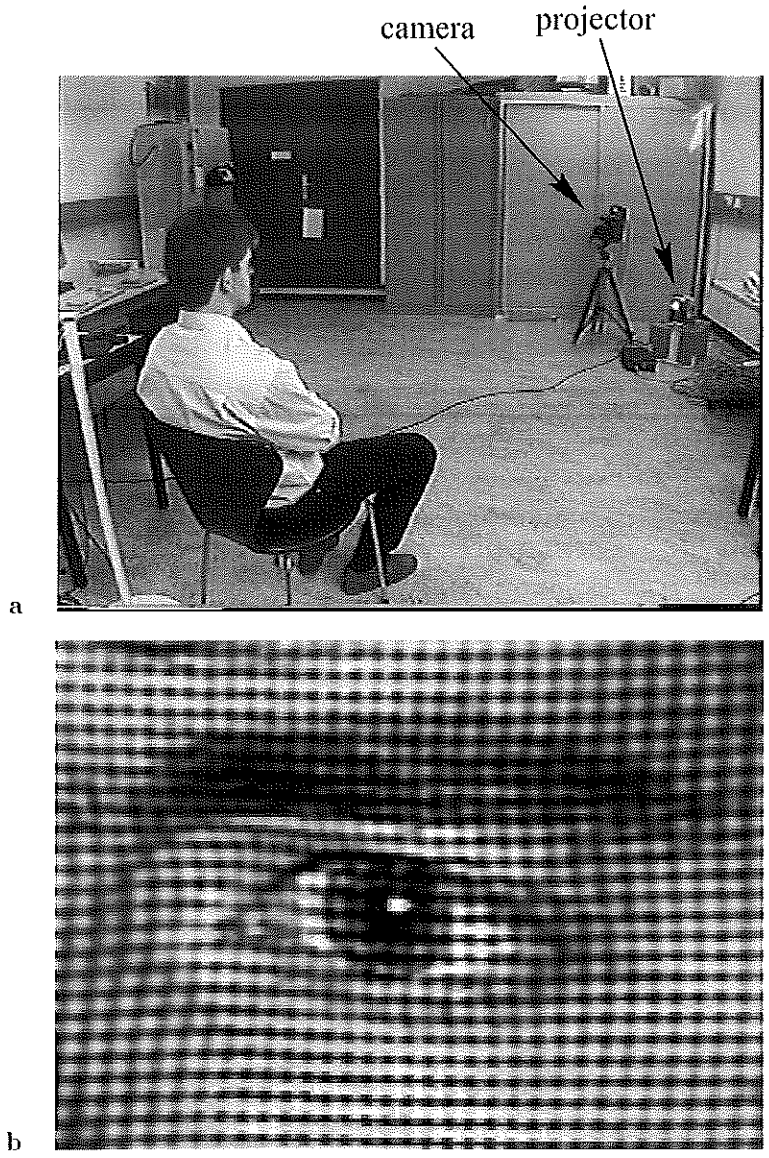
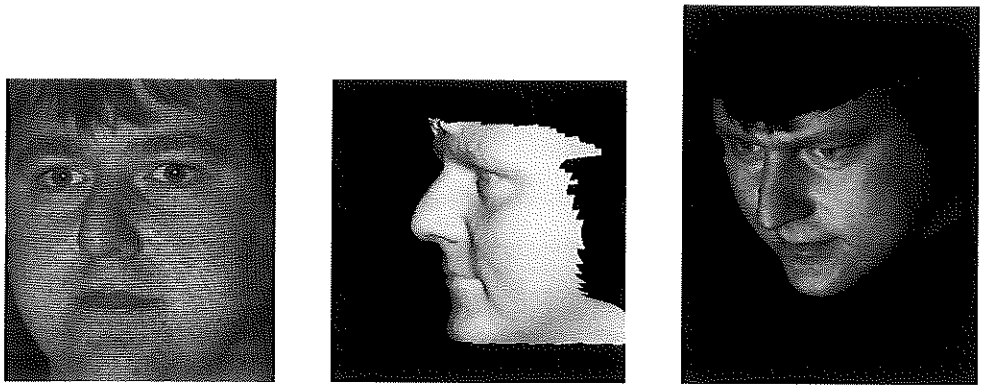


Figure 2: **a:** *Required hardware: a camera, a projector and a computer. A regular square pattern is projected on the scene (here the face of the person sitting on the chair).* **b:** *Detail of the face, with the projected pattern.*



a b c

Figure 3: a: *Original image of a face;* b: *A profile view of the reconstruction;* c: *Texture mapped reconstruction. All results were obtained from the image in* a.

The one-shot nature of this system, i.e. the extraction of the 3-D data from a single image, makes it possible to take a video sequence of a face, and reconstruct every frame of it. The result is a dynamic, 3-D reconstruction. Such dynamic 3-D sequences could be of particular interest for the study of facial expressions, the extraction of visemes, etc. The one-shot nature of this system also allows to extract the 3-D shape of objects while in motion, a consideration that might be of particular importance for visual inspection along production lines. The capture of dynamic 3-D is still quite rare, but a few other systems are around. Kanade and coworkers have demonstrated the extraction of 3-D, whole body dynamics with a passive system with about 50 cameras [21]. Nayar and coworkers [25, 15] used shape from defocus. Multiple aligned cameras take images of the same scene through a system of half-translucent mirrors. Their system also uses a simple, regular illumination pattern. To the best of our knowledge, the system doesn't extract surface texture, but such an extension would seem feasible.

The system used here is also not alone in its one-shot operation. Hall *et al.* [7] developed a grid-pattern method for extracting sparse range images for simple shapes, where the identification is based on the interactive labeling of some line intersections. Vuori and Smith [24] based their line identification on the restriction that the objects have a maximal height and lines will fall in predictable stripes of the image. Therefore, the grid has to remain quite sparse. Blake *et al.* [3] use a calibrated stereo setup to observe two sets of subsequently projected lines. By having the lines intersect the epipolar lines obliquely, these intersections help narrow down the possible line identity. Again, lines must not be positioned too close in order for this strategy to work. As a result, the reconstructions obtained with these systems all have

a rather low resolution.
Other systems use a "single" image of denser grids, but with some form of line coding. Boyer and Kak [4] developed a light striping concept based on colour-coding. Vuylsteke and Oosterlinck [23] use a binary coding scheme where each line neighbourhood has its own signature. Maruyama and Abe [14] introduce random gaps in the line pattern which allows to uniquely identify the lines. An exception is the work by Chia *et al.*[6]. The authors assume orthographic projection of a pattern combined with sufficient perspective distorsions in the camera image. The latter is necessary to identify the lines in the pattern. They also assume that the intrinsic camera parameters are known. We believe that the strategy behind the system proposed here is more robust, because it doesn't depend on subtle effects of perspective deformations. A unique feature is that no attempt is made to identify the lines in absolute terms. Only the relative positions of the lines in the pattern is used.

More details about the system can be found elsewhere [20].

3 Illustration of 3-D face reconstructions

This section illustrates some of the typical results obtained with the system.

Fig. 4 shows 3-D reconstructions for 5 faces. The left column shows the input images. The other three columns show the 3-D reconstructions from different viewpoints. Note the gaps in some of the views (black areas): parts of the faces that were not visible to the camera could not be reconstructed. Taking multiple views can remedy this problem. This is illustrated in fig. 5 where 2 views are taken of a face and the resulting reconstructions have been brought in registration, as shown on the right. In this case, part of the nose is hidden for each of the two camera positions. The combined reconstruction (fig. 5, image on the right) shows that the two reconstructions fit well. Bright areas are those for which data are available in the two reconstructions, dark areas correspond to data extracted from the first view exclusively. The registration was carried out using the ICP algorithm [5]. The holes in the reconstructions show that the system successfully handles depth discontinuities. A bigger problem is hair. Typically, hair is not captured well and these parts are also left open by the system.

In the previous illustration, the strategy has been to use as few images as possible for the complete reconstruction of a face. The resulting 3-D descriptions consist of approximately 8000-10000 bilinear patches. If this resolution is too low, one can project a finer grid and either take a higher resolution camera (e.g. one of the latest digital photo cameras), or zoom in and compile the 3-D model out of smaller pieces. The latter is possible by only moving the camera, and without additional calibration. On the statue, a grid was projected of 600×600 lines. This is too many to even be visible in a single image. Hence, one can 'scan' the scene by taking a series of more

Figure 4: *3-D reconstructions of 5 faces. The left column shows the input images.*

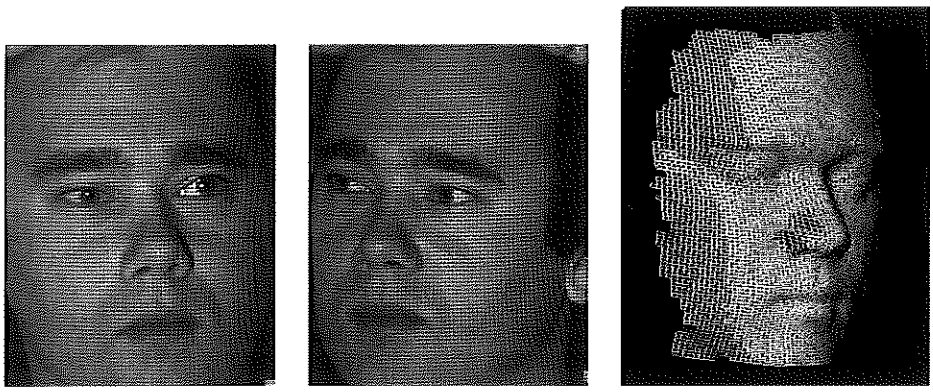


Figure 5: *The same face from two different camera viewpoints. The image on the right shows the superposition after registration of the two reconstructions.*

detailed images. While doing so, the position of the projector is kept fixed. Only the camera is moved around, taking images of the detailed patches. The setup is calibrated only once, best for the camera position that corresponds to the most central patch. The patches have to overlap, in order to support registration afterwards. A series of such images is shown in fig. 6. For every image a reconstruction is made of the corresponding surface patch. Only the reconstruction of the patch for which the calibration was carried out will be 'correct', however. The others have – by approximation – an affine skew. Therefore, a registration step was implemented that allows them to deform affinely while being matched. Fig. 7 shows the resulting mosaic of matched patches. As all the patches are based on a single grid of lines, it is easy to integrate them into a single surface description. As to the texture, this can be extracted as a weighted sum of the overlapping image textures. Fig. 8 show the result for the statue. As these results show, because of the high resolution one can zoom in quite a bit on parts of this structure while keeping a realistic impression. Of course, the reconstruction is still only a partial one. Such large, but high resolution parts can in a second pass be registered as well, to form a complete reconstruction.

In the previous examples, the face to be capture remained still. As mentioned before, the system holds special promise for the extraction of dynamic 3-D. Fig. 9 shows a few frames of a video taken for the 3-D reconstruction of facial expressions. The grid was projected onto the face all the time. For every frame of the video, a 3-D reconstruction was made and the skin texture was extracted. Hence, not only shape but also texture dynamics are captured, which is important e.g. to include the blinking of the eyes. Fig. 10 shows the results for the frames selected in fig. 9. The reconstructions are viewed from three different directions. Armed with such dynamic 3-D data,

Figure 6: *A series of detailed image of a Greek statue. The corresponding patches overlap.*

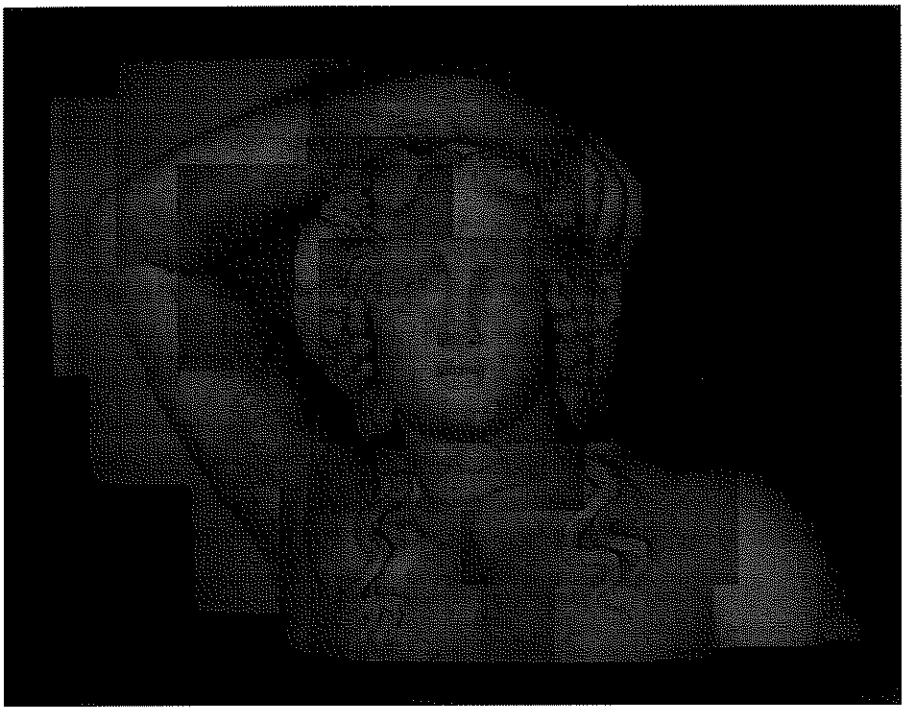


Figure 7: *The mosaic of patches after registration.*

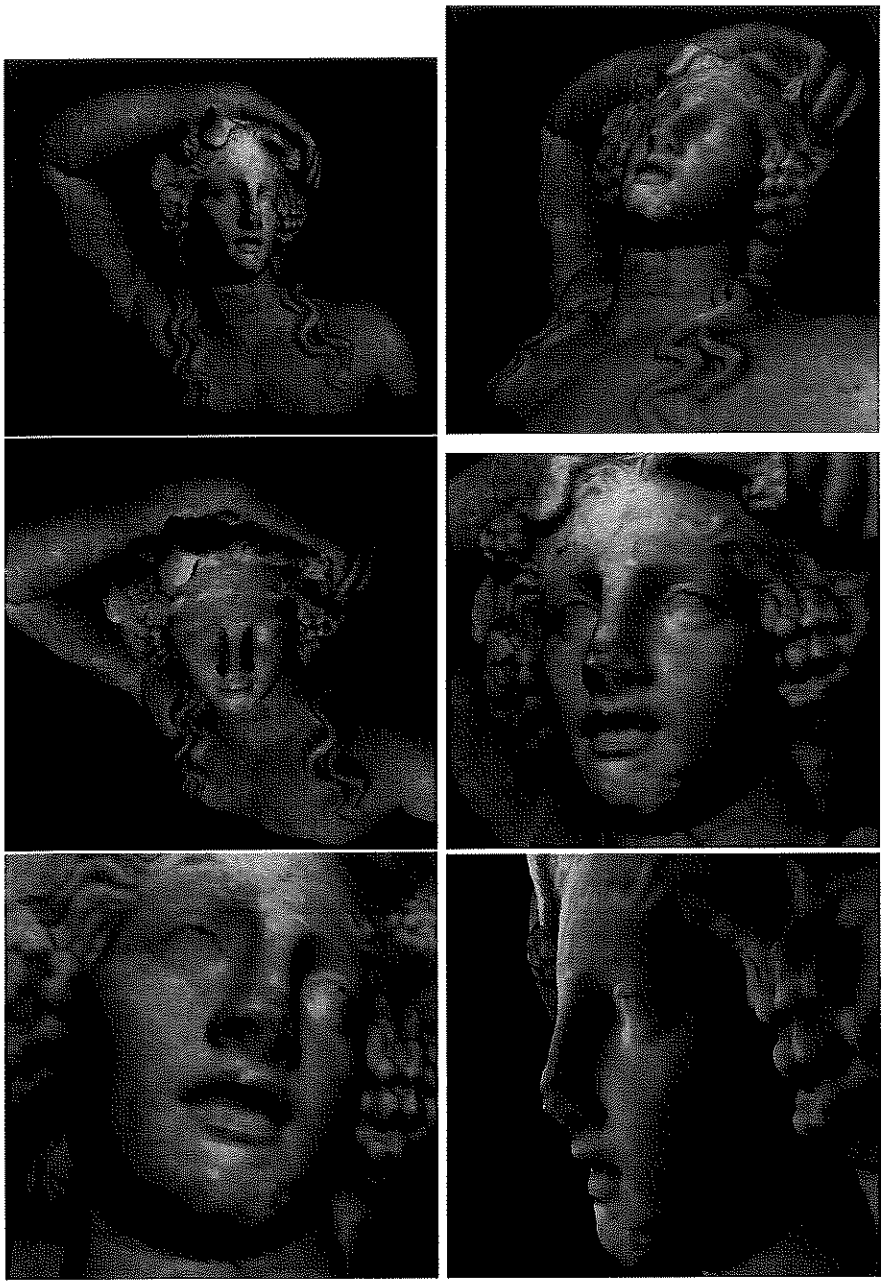


Figure 8: *Six views of the reconstructed statue. The reconstruction has been built from different patches, as explained in the text.*

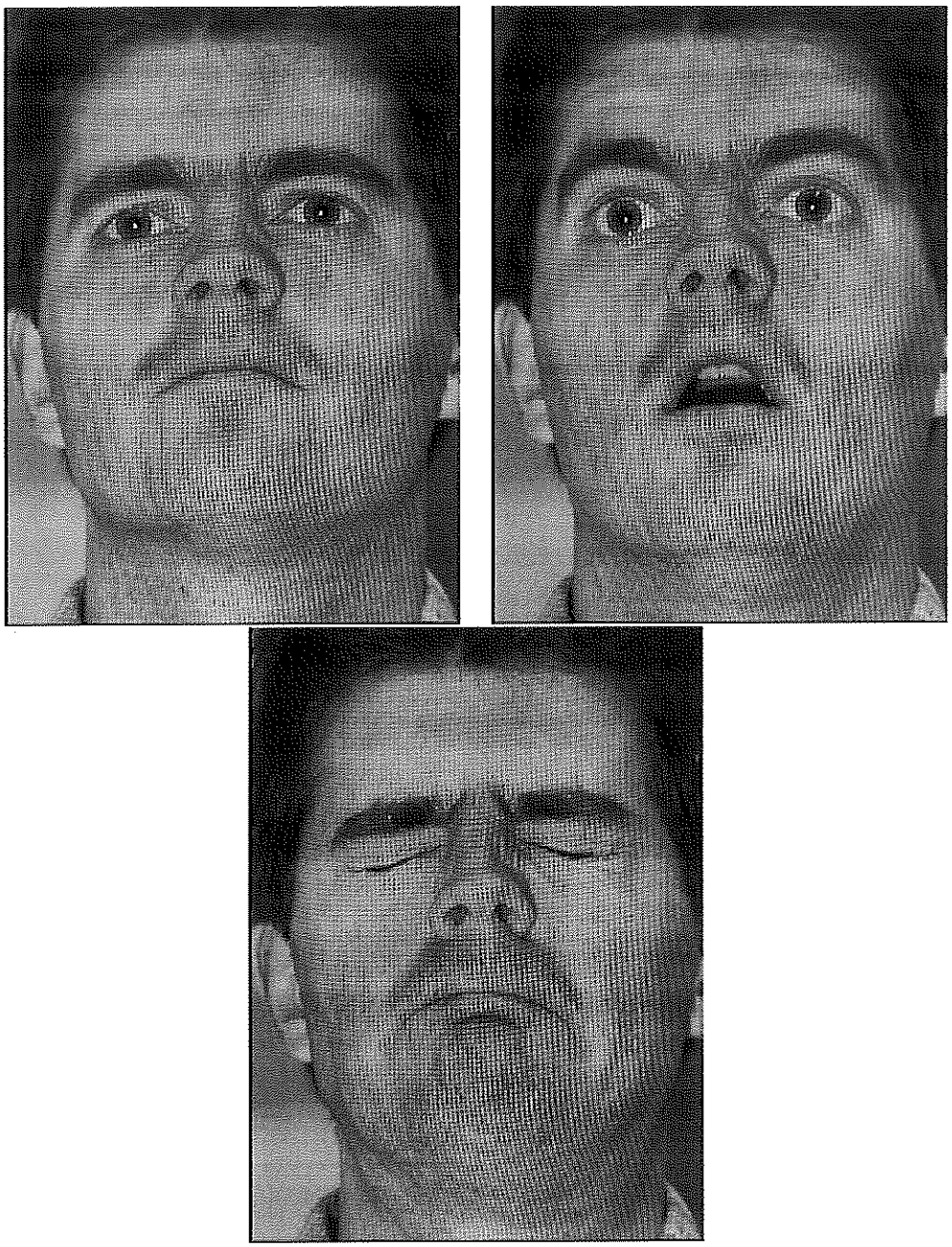


Figure 9: *Three frames out of a video sequence showing a gesturing face*

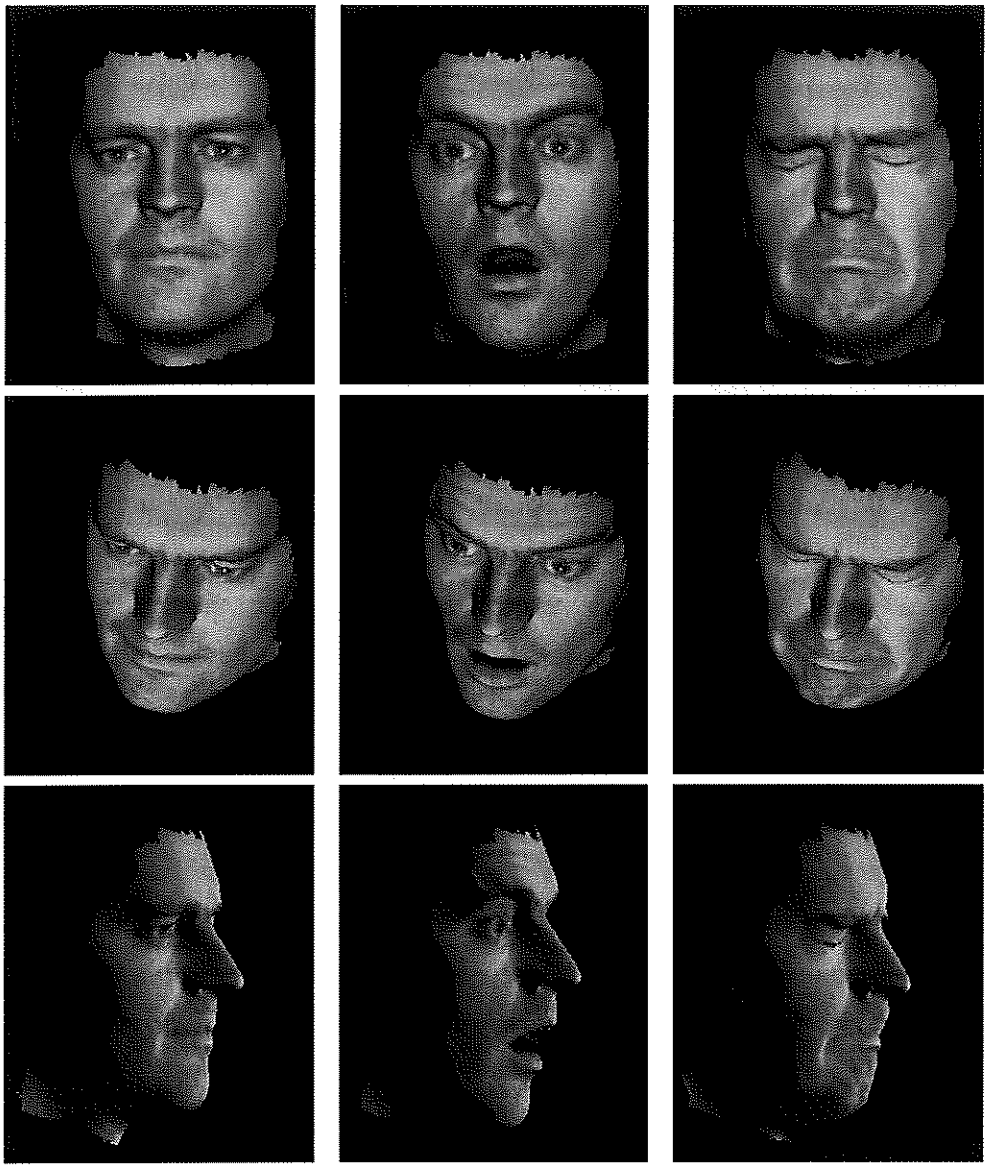


Figure 10: *Texture-mapped 3D reconstructions for the frames of fig. 9 shown from 3 different viewpoints.*

it is tempting to use the observed changes directly as a means for driving face animation suites. Rather than going through the painstaking process of modelling the skull, muscles, and skin, the 3-D optical flow of the data could be calculated, the typical motions could be extracted for different facial expressions over several individuals, and applied to the corresponding points on other faces, much in line with the approach pioneered by Thomas Vetter (cfr. same volume).

4 Forensic applications of 3-D face models

Three-dimensional face data can play a useful role in the identification of criminals. The comparision of surveillance video data and mugshots of potential offenders or suspects often is a difficult task. Surveillance cameras typically look down upon the scene, whereas mugshots usually are frontal or profile pictures. Similarly, the illumination conditions can be quite different. Sometimes the police forces will take a suspect to the scene of the crime and images can be taken with at least the same camera and from a similar viewpoint. Such procedure is time consuming, expensive and not always without risk, however.

The availability of a 3-D head model for a suspect can alleviate such problems considerably. It then becomes possible to depict a suspect's head in a similar 3-D position and to emulate the lighting that was in place at the time of the crime. It is then much easier to make direct comparisons with the surveillance data. It becomes e.g. possible to overlay a number of facial features and to check whether the rest of the faces fall in registration.

The possibility to show a 3-D face model from different relative viewpoints, including those of the original surveillance cameras, is obvious. The emulation of changing illumination is not so obvious at this point. The texture that is mapped onto the faces is obtained from the image that is used for the extraction of the 3-D shape. Image based texture does not yield the true surface reflectance, however, and this is needed to simulate changes in illumination. Therefore, from the image texture, the illumination at the time of data capture has to be decoupled from the surface reflectance characteristics. In general, this is a very difficult problem. Fortunately, the situation here is less complex, because the angle of the incoming (projected) light is calculated explicitly when the system is calibrated. Armed with the 3-D shape and the knowledge of where the light is coming from, reflectance modelling becomes much easier. Nevertheless, some problems remain, because the derivation of a BRDF would require a sufficient number of samples. To that end, assumptions have to be made on parts of the face having similar reflectance characteristics. Moreover, with the current setup the angle between viewing and projection is always small. It is also a bit naive to assume that all light is coming from a single direction, i.e. from the projector. In a typical room,

there will be ambient light. So far, our work in this area has been restricted to the modelling of the latter effect.

Our preliminary experiments start with the assumption that Lambertian reflection is a good model for most of the face. This might seem far-fetched in view of the observations that have been made for diffuse reflection from real surfaces [16]. The angle between the rays of projection and viewing is quite small (typically 10^o or less) and under these conditions the Lambertian model can be expected to apply rather well [28]. Assuming the face surface is Lambertian,

$$A \sim \frac{I}{\cos(\alpha)} \ ,$$

with A the albedo, I the image intensity, and α the angle between the surface normal and the incoming light. Using this model, the resulting albedo is shown in fig. 11b. One would expect to find values that are more or less

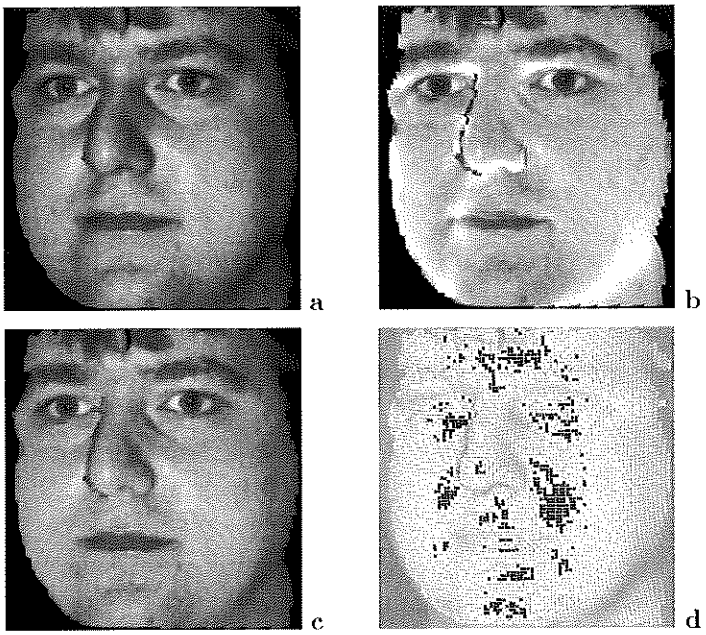


Figure 11: **a:** *Original texture mapped surface.* **b:** *'Albedo' for Lambertian surface.* **c:** *'Albedo' with the proposed lighting model.* **d:** *Black areas denote places where specular reflections may occur.*

constant over the face, but under the jaw substantially higher values are found. This clearly is not correct, but it is in line with the expectations about diffusely reflecting surfaces. It is typical that the Lambertian model yields albedos that are too high where the normal on the surface deviates strongly from the incoming light direction [27]. A second factor that could contribute

is ambient light. This assumption was tested by refining the illumination model. In addition to the incoming, directed light, it was assumed that there was a constant, omnidirectional ambient component (e.g. reflections from the walls etc.). In that case

$$A \sim \frac{I}{f(\alpha)}, \quad \text{with} \quad f(\alpha) = 1 - \gamma + \gamma.\cos(\alpha) \ ,$$

with γ a number between 0 and 1 that determines the relative weight of the directional and ambient light components. In a little experiment, the value of γ was chosen as to minimize the variations in 'albedo' over the face. For the example image, the optimal value came out to be $\gamma = 0.3$ and the resulting albedo is given in fig. 11c. As can be seen, there are still substantial variations in the value of the albedo. The brighter regions correspond to specularities, which would require additional care. The regions underneath the chin and the jaw are darker now. This again is a deviation from the expectations.

It therefore seems necessary to modify the Lambertian model itself, both by refining it according to the refined models for diffuse reflection [16, 27, 28] and by adding a specular component. The places where specular reflection might occur can be predicted rather well. The strongest specularities will be caused by the directional light and its incident direction is known with respect to the surface normal and the viewpoint of the camera. Fig. 11d shows the places where the mirror conditions for incoming and outgoing light are satisfied (black dots). It is interesting to note that these positions can be found without a 3-D reconstruction. They correspond to positions where the grid looks square in the camera image. This is caused by the fact that enlarging and foreshortening effects of projection and viewing resp. cancel out, due to the mirror configurations of projecting and viewing rays.

An example of a virtual change of illumination, based on the albedo resulting from the mixed lighting model, is shown in fig. 12

5 Facial feature extraction in 3-D

Most current work on facial feature extraction takes video sequences as input. It has proved not so straightforward to achieve good robustness from such images. It can e.g. be very difficult to extract the lips, certainly if the images contain a complete face and the resolution of the mouth is not so high. In order to build stable lip detectors and trackers researchers usually had to control the viewpoint, to zoom in on the mouth area or to use lipstick to increase contrast (see e.g [11, 12, 2] for state-of-the-art contributions).

For one thing, the availability of 3-D data can help to reduce the influence of the viewpoint, through the use of viewpoint invariant geometrical features. What we propose is a kind of syntactical approach to facial feature extraction, much in the tradition of the face recognition literature but based on 3-D

Figure 12: *A simulated change in illumination (right image) on the basis of the original view (left image).*

.

data. The first step is to detect the nose as the point where both principal curvatures κ_1 and κ_2 are large. The nose tip can thus be defined as the point where one finds

$$\max(\sum_{W_x} min(\kappa_1, \kappa_2)) \ .$$

in some window with width W around it. The position of the nose tip gives a first indication of the location of the mouth area. Each lip can be modeled as a long stretched region with one large and one small principle curvature. Furthermore it is assumed that both lips have the same extent, lie almost parallel to eachother, and have approximately the same curvature. Both lips are determined simultaneously as

$$\max(\sum_{lip\ region} min(\kappa_1^{upper\ lip}, \kappa_1^{lower\ lip})) \ .$$

with κ_1 the largest principal curvature. Similar methods can be applied to find the eyebrows and the chin.

The results of the nose and mouth extraction procedures are illustrated on a series of images. A video sequence was taken of a talking face. Throughout, the reconstruction grid was projected onto the face. Fig. 13 shows 3 frames from the talking head sequence. Fig. 14 shows a part of the original images, with the position of the nose indicated with a point and the lips as two line segments. Both the nose and lips have been detected at a resolution given by the squares of the pattern. Each frame was treated separately, so no tracking was used to improve the results and the results therefore illustrate the quality that can be expected from a single view. As can be seen, the position of the nose point is stable and the lip lines nicely stick to the upper and lower parts of the upper and lower lips, also when the teeth become visible.

Figure 13: *Three frames of a talking head video sequence.*

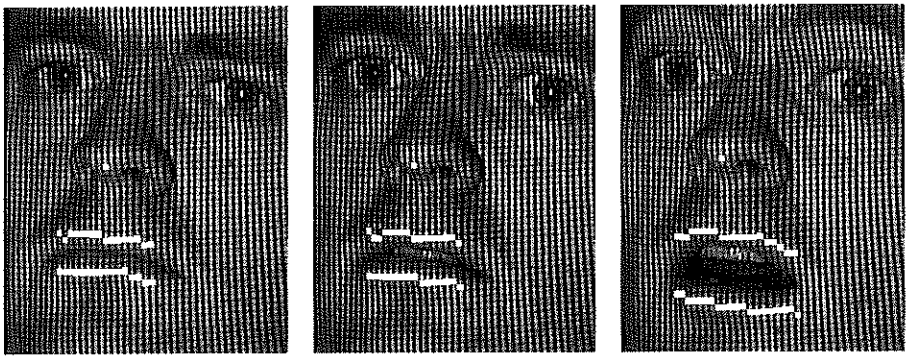


Figure 14: *Detected nose and lips, based on surface curvature. Results for every frame were obtained independently.*

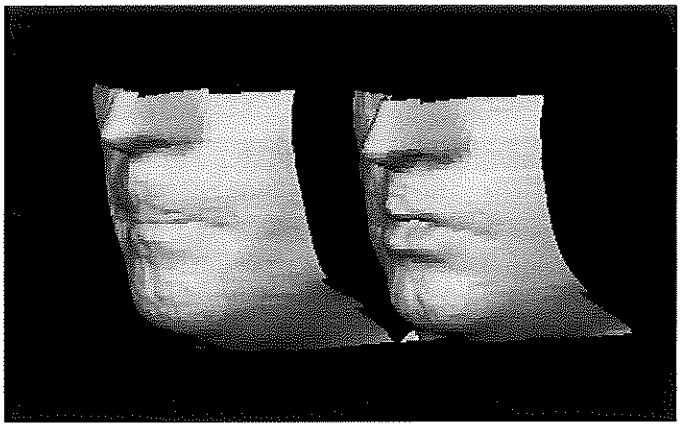


Figure 15: *3-D mouth shapes (visemes) for /m/ (left) and /n/ (right).*

Facial feature tracking is useful for several applications. One is the creation of intelligent human-machine interfaces. The machine can be made to look at the facial expressions an adapt its behaviour accordingly. Another example is 'visual speech'. In noisy environments, the performance of speech recognition systems can be improved by simultaneously looking at the mouth. Rather than apply speech recognition or lip reading in isolation, both can be used to complement each other. As an example, /m/ and /n/ sound very similar, but can be distinguished quite easily by their 'visemes', i.e. the shape of the mouth area. This is illustrated in fig. 15. A further example where facial features have a role to play is in the world of virtual actors, special effects, and related issues in the postproduction industry for broadcasting and the movies. Fig. 16 shows six frames of a small movie. The face was put through a virtual ordeal, being submerged, deformed, tossed and turned. The input was the gesturing face sequence already shown in fig. 10. In order to implement this demo, facial features like the nose and chin were automatically detected and deformed. Note how also the orientation and illumination constantly change. Achieving the same level of realism using graphics-based animation would take quite an effort.

6 Conclusions and future work

The paper focused on an active technique to generate 3-D models of faces. These were used for the extraction of basic textural and geometric features. The proposed 3-D acquisition method also allows capturing dynamic 3-D data, rendering it especially useful for the analysis and reconstruction of facial expressions and visemes. In summary, the main characteristics of the proposed acquisition method are:

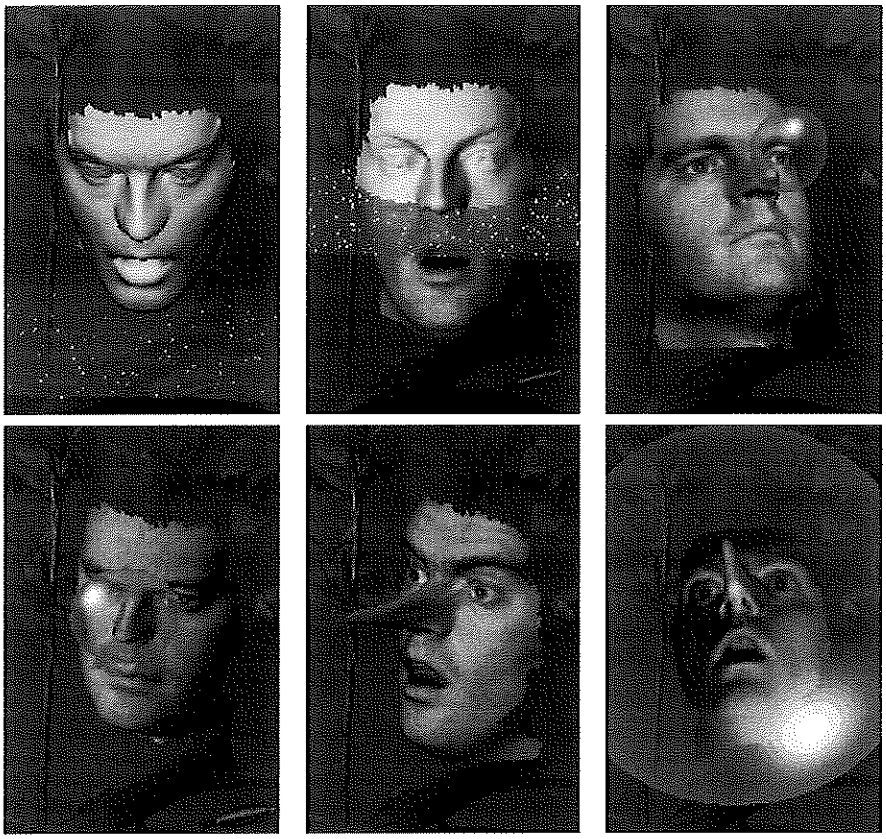


Figure 16: *A number of frames out of the VR demo sequence.*

1. the required hardware is minimal – a slide projector, a camera and a computer – and hence the system is cheap;

2. the calibration is simple, requiring no exotic calibration objects or patterns, and making the system easy to transport;

3. the acquisition time for the necessary input is that one of a single image at standard video frame rate, therefore head motion is not a problem and can be captured;

4. the spatial resolution is sufficiently high for most facial feature extraction tasks;

5. the output is a mesh of connected points, rather than a mere cloud of points, hence the topology of the surface is known;

6. it extracts realistic surface texture, in colour if required;

Planned work encompasses the following aspects:

- Speeding up the system. Currently it takes a few minutes for the extraction of the 3-D shape and texture for a single frame. A first and self-evident improvement would be to exploit the temporal continuity that exists between the shapes and textures of subsequent frames. All dynamic results shown in this paper have been obtained by carrying out the reconstruction of each frame from scratch. In fact, the stability of the reconstructions when viewed as a video testifies to the precision of the method.

- Refining the skin reflectance model. Given the rather detailed information from the image textured 3-D models, it seems worthwhile spending more time on the derivation of refined skin reflectance models.

- Combining photometric and geometric cues. For the extraction of the facial features only geometrical cues were used. Combining these with texture cues looks like a promising avenue.

- Extending the applications. We have plans to work on several applications of 3-D faces. One is building tools that assist the police forces with the identification of criminals. Another is the animation of virtual actors based on a real actor's performance, or visual speech extraction.

Acknowledgements: M.P. gratefully acknowledges a postdoctoral research grant from the Flemish Institute for the advancement of Science and Technology in Industry (IWT). The authors also gratefully acknowledge support from Esprit-LTR project 'Improofs' for the reported skin reflectance work.

References

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

[2] A. Blake, R. Curwen, and A. Zisserman, A Framework for Sptio-Temporal Control in the Tracking of Visual Contours, Int. J. Comp. Vision, Vol 11.2, pp. 127-145, 1993.

[3] A. Blake, D. McCowen, H. R. Lo, and P. J. Lindsey, Trinocular Active Range-Sensing, IEEE PAMI 15(5), pp. 477-483, 1993.

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

[5] Y. Chen and G. Medioni, Object Modeling by Registration of Multiple Range Images. Proc. Int. Conf. on Robotics and Automation, Sacramento CA, pp. 2724-2729, 1991

[6] T. Chia, Z. Chen and C. Yueh, Curved Surface Reconstruction using a Simple Structured Light Method, Proc. Internat. Conf. Pattern Recognition, Vienna, Vol. A, pp. 844-848, 1996

[7] Hall, E.L., Measuring curved surface for robot vision, IEEE computer, Vol. 15, No. 12, p.42-54

[8] G. Hu and G. Stockman, 3-D surface solution using structured light and constraint propagation, IEEE Trans. PAMI, Vol. 11, No. 4, pp. 390-402, 1989

[9] Jarvis, A persepctive on range finding techniques for computer vision, IEEE Trans. on PAMI, Vol. 5, No 2, March 83, p.122-139

[10] A. Lanitis, N.A. Thacker, and S.W. Beet, A Unified Approach to Coding and Interpreting Face Images, Proc. ICCV, 1995.

[11] M. Kass, A. Witkin, and D. Terzopoulos, Snakes: active contour models, Int. J. Comp. Vision, pp. 321-331, 1988.

[12] R. Kaucic, B. Dalton, and A. Blake, Real-time Lip Tracking for Audio-Visual Speech Recognition Applications, Proc. ECCV, Vol II, pp. 376-387, 1996.

[13] Y. Lee, D. Terzopoulos, and K. Waters, Realistic modeling for facial animation, SIGGRAPH, pp. 55-62, 1995

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

[15] S.K. Nayar, M. Watanabe, and M. Noguchi Real-time focus range sensor, IEEE Trans. PAMI, Vol.18, No.12, pp. 1186-1197, 1996.

[16] M. Oren and S. Nayar, Generalization of the Lambertian model and implications for machine vision, Int. Journal of Computer Vision, 14(3), pp. 227-252, 1995

[17] M.A. Turk and A.P. Pentland, Face Recognition Using Eigenfaces, Proc. CVPR, pp. 586-591, 1991. 1990.

[18] M. Proesmans, L. Van Gool, and A. Oosterlinck, Determination of optical flow and its discontinuities using non-linear diffusion, Third European Conf. on Computer Vision, Stockholm, pp. 295-304, may 1994 ed.

[19] M. Proesmans, L. Van Gool and A. Oosterlinck, One-shot active range acquisition, Proc. Internat. Conf. Pattern Recognition, Vienna, Vol. C, pp. 336-340, 1996

[20] M. Proesmans and L. Van Gool, One-shot active 3D shape acquisition. Multisensor Fusion and Integration, to be held in Washington DC, Dec.'96.

[21] P. Rander, P. Narayanan, and T. Kanade, Revovery of dynamic scene structure from multiple image sequences, Proc. Int. Conf. Multisensor Fusion and Integration of Intell. Systems, pp. 305-312, 1996

[22] Rioux M., Laser range finder based upon synchronous scanners. Applied Optics 23(21), p.3837-3844, 1984.

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

[24] Vuori T. A. and Smith C.L., Three dimensional imaging system with structered lighting and practical constraints, Journal of Electronic Imaging 6(1), pp. 140-144, 1997.

[25] M. Watanabe and S. Nayar, Telecentric optics for computational vision, Proc. European Conf. Computer Vision, Vol. II, pp. 439-451, 1996.

[26] Will P.M. and Pennington K.S. Grid coding: a novel technique for image processing, Proceedings IEEE, 60(6), pp. 669-680, 1972.

[27] L. Wolff, On the relative brightness of specular and diffuse reflection, Proc. CVPR, pp. 369-376, 1994.

[28] L. Wolff, Generalizing Lambert's Law For Smooth Surfaces, Proc. ECCV, pp. 40-53, 1996.