3D modeling for communications

Luc Van Gool, Filip Defoort, Reinhard Koch, Marc Pollefeys, Marc Proesmans, and Maarten Vergauwen
ESAT-PSI, Katholieke Universiteit Leuven
Luc.VanGool@esat.kuleuven.ac.be

Abstract

The media and communications providers share an increasing interest in 3D models of people, objects, and scenes. The paper focuses on features that 3D acquisition systems ought to have in order to optimally serve these markets, where emphasis is on realistic visualisation. It is argued that 3D acquisition techniques developed for traditional applications like visual inspection not necessarily are the best option. Techniques should be developed that are dedicated to visualisation-specific requirements. This is exemplified with two systems that have been developed recently. One takes uncalibrated video data as input from which it generates a 3D model. A second system projects a grid of lines and gets dense 3D from a single image. This system needs some calibration, but the corresponding procedure is extremely simple. It can also be used to capture detailed, 3D scene dynamics.

1 Shifts in 3D

Traditionally, 3D acquisition technology has been developed for visual inspection and robot guidance. These two domains have for a long time also been the two major driving forces behind computer vision in general. Nowadays however, telecommunications and other domains with a need for realistic visualisation – and this encompasses the media in general – have become the primary motor behind developments in computer vision.

Quite a few of these newer developments involve solutions to problems nobody had really been dealing with before. Think of database retrieval, visual speech, or the recognition of emotions from facial expressions. By contrast, 3D modeling had quite a history already and consequently it was tempting to use existing technology for the creation of 3D models in these new application areas. As a result, laser scanners are still the 3D acquisition technology of choice for many.

Yet, the novel applications of 3D acquisition technology come with requirements of their own:

- For one thing, the extraction of surface texture is a must. Visualisation of untextured 3D surfaces is anything but unacceptable for most applications in the (tele)communications arena. For instance, people find it even surprisingly difficult to recognize a person's face when shown a detailed 3D but untextured representation.
- The absolute scale of shapes is of much less importance. Their images may be shown at quite different scales anyway,
- There is a larger variety of object sizes to be handled. Whereas the traditional industrial setting would allow for a predefined and probably rather narrow range of sizes, applications that require 3D modelling for visualisation typically deal with substantial variation of sizes. Think of the different objects used in special effects for the movies. In 'Twister' we saw spiral anything from bricks to large fuel trucks into the sky.
- There is a tremendous advantage to be gained if dynamic 3D can be extracted, ie. 3D shapes together with their changes over time. This is a dream come true for people in animation, who now have to go to great length in order to extract natural motions from observed markers in performance animation or through the tedious modeling of muscle and skin dynamics.
- As the users will probably be postproduction people, artists, or even participants in games rather than engineers from the quality control department whose primary job it is to operate the 3D acquisition system, these new 3D acquisition systems should be easier to use and easier to install. No mindboggling calibration procedures here... This is all the more important as for these visualisation oriented markets it is more often necessary to freely move around the whole apparatus.

Traditional 3D acquisition systems do not quite match this shopping list. Laser scanners and stereo systems are most widespread.
The former have a working volume that is restricted primarily by the need for precise mechanical scanning, cannot

capture texture without additional provisions and/or manual tinkering, and have acquisition times that are too long to allow for substantial object motion during scanning.
Stereo systems are more flexible in terms of the working volume, they capture the texture and shape without additional problems of aligning the two, but require careful calibration when used in the traditional triangulation framework. They can in principle extract 3D object motion. In general, the visual quality of the 3D models that are obtained from a mere pair or triple of images is insufficient, as point correspondences are difficult to find precisely even under perfect calibration. Moreover, that calibration typically is a tedious process.

The vision group at K.U.Leuven is working towards more flexible and robust versions of these techniques. Here we concisely describe two of these attempts. One could be considered a counterpart for the laser scanner in that it also projects a special light pattern. The second technique is more stereo-like in that it takes multiple, plain images as input. In both cases 3D shapes can be acquired very easily. Emphasis will here be on the first technique, as the second one is the subject of a separate presentation in these proceedings (Koch *et al.*). The second will therefore only be described very succinctly, in the next section.

2 Shape-from-video

The second technique starts from multiple images, e.g. a video sequence. In contrast to traditional shape-from-motion or stereo approaches, the relative camera positions for subsequent views are not known and neither are the internal camera parameters (focal length, pixel aspect ratio, etc.). Hence, this technique can start from very general data, e.g. old video footage of a building that no longer exists, amateur video data that were not taken with 3D reconstruction in mind, etc. There are a few limitations, but these are relatively minor. Typically, the baseline between the images has to be small on a pairwise basis, i.e. for every image there is another one that has been taken from a rather similar viewpoint. The reason is that finding correspondences between different views should not be made too difficult for the computer. For the rest, the approach hardly imposes any constraints on the camera motion, except that one has to avoid some particular, degenerate motions like pure translation throughout the image sequence.

The method is based on the automatic tracking of image features over the different views. This is done in stages. First, a 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 and trilinear constraints). These constraints support the correspondence search for a wider set of features and in the limit, for a dense, i.e. point-wise, field of disparities between the images. The limited set of corner correspondences also yields the camera projection matrices that are necesssary for 3D reconstruction. In general, to arrive at metric structure – i.e. to undo any remaining projective skew from the 3D reconstruction – the camera internal 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. Rather than dwelling further on this here (as mentioned, this work is the subject of a companion paper), consider the example of fig. 1. 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. 2 shows 3 views of the 3D reconstruction, from different viewpoints than the input images.

Figure 1. Two of 6 images of the Indian temple.

The development of such technique was based on the contributions of a whole community of vision researchers. References are given in the companion paper. The remainder of the paper focuses on the active technique.

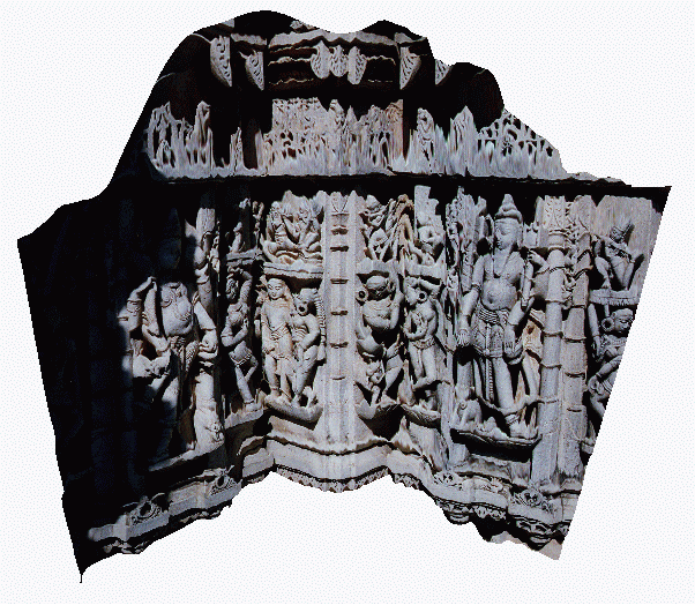

Figure 2. Three views of the reconstruction obtained for the Indian temple.

3 Active, one-shot 3D acquisition

The 'passive' technique outlined in the previous section cannot deal with untextured parts of a scene. Yet, from a graphics or telecommunications perspective, there are very important applications where this problem pops up. Think of faces, which are only slightly textured at places. For such surfaces it is difficult to get the precise correspondences between points in different views. Therefore, there remains an important role for 'active' 3D acquisition systems. Active systems extract 3D from the displacements/deformations of a projected pattern when observed from a viewpoint different from the position of the projector. These systems achieve higher precision through the projection of the illumination pattern, which provides the necessary cues to solve the correspondence search. With the computer vision group in Leuven we use a *dense*, square grid of lines..

Also this work is related to a substantial body of literature. A classic reference on range-imaging methods is the survey of Jarvis [4]. Besl [1] gives a broad overview of available active optical range sensors and gives a quantitative performance comparison. Typically such methods have relied on the projection of (single) points or (single) lines to scan the scene and 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 to aid in identifying the lines. Such constraints could e.g. be a maximal depth range for the objects. In particular, the identification of the lines can be jeopardized in the presence of depth discontinuities or when not all the lines are visible. For a discussion on the problem of line labeling and identification, see Hu and Stockman [3].

Still with the aim of identifying the individual lines of a grid, an option is to add a code. Boyer and Kak [2] developed a light striping concept based on colour-coding. Vuylsteke and Oosterlinck [6] used a binary coding scheme where each line neighbourhood has its own signature. Maruyama and Abe [5] introduced random gaps in the line pattern which allows to uniquely identify the lines. The need to extract a code adds some vulnerability to these systems and also tends to keep the maximal number of grid lines rather low.

With the proposed technique, dense grids can be projected and hence high-quality 3D reconstructions can be made from single images. The crux of the matter is to avoid the need for line identification altogether.

3.1 Outline of shape extraction

The proposed method is based on the projection of a single pattern, which is a square grid of lines. No code has to

be included in this pattern. The object to be reconstructed is observed by a single camera. Fig. 3 shows the setup and an example image from which 3D information is extracted. Because there are many lines and these are put closely together, the individual identification of the lines is virtually impossible. However, relative 3D positions of points can be obtained directly from relative positions of grid line intersections in the image. No detour via line identification or the integration of differential surface properties like curvature of surface normals is necessary. This is the basic novelty underlying the system, by which it is possible to use very dense grids, with lines separated by not more than a few pixels in the image.

In contrast to the passive approach of the previous section, this one needs a little bit of calibration. This step has to supply the relative orientations of the projector and the camera. From the user's point of view the procedure is extremely simple. It suffices to show the system a scene dominated by two planes that subtend a known angle. Typing the angle is all it takes to calibrate the system.

Only those parts can be reconstructed that are visible from the camera and receive light from the projector. The surface area satisfying these constraints is increased by putting the projector and camera closer to eachother. This also makes it possible to model narrow cavities in the shape, that yield self-occlusion problems with larger angles. A caveat is that the precision of the 3D reconstruction also deteriorates when the directions of projection and viewing come closer. Nevertheless, precision can be kept acceptable with angles as small as 10^o or even smaller. This is also the range of angles used in the examples of section 3.2. The angle can be made so small precisely because the shape extraction goes directly for all the intersections of the 2D projection pattern instead of dealing with individual points or lines (as with traditional laser scanners).

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 being projected. 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.

3.2 Examples

Fig. 4 shows the image of a dwarf figure with the grid projected on it. Fig. 5 shows four views of the reconstructions obtained from this single image. The reconstruction is of good quality even close to the object boundaries in the image. This is an advantage of using a small angle between the directions of projection and viewing. The size of surface patches that can be reconstructed from a single camera

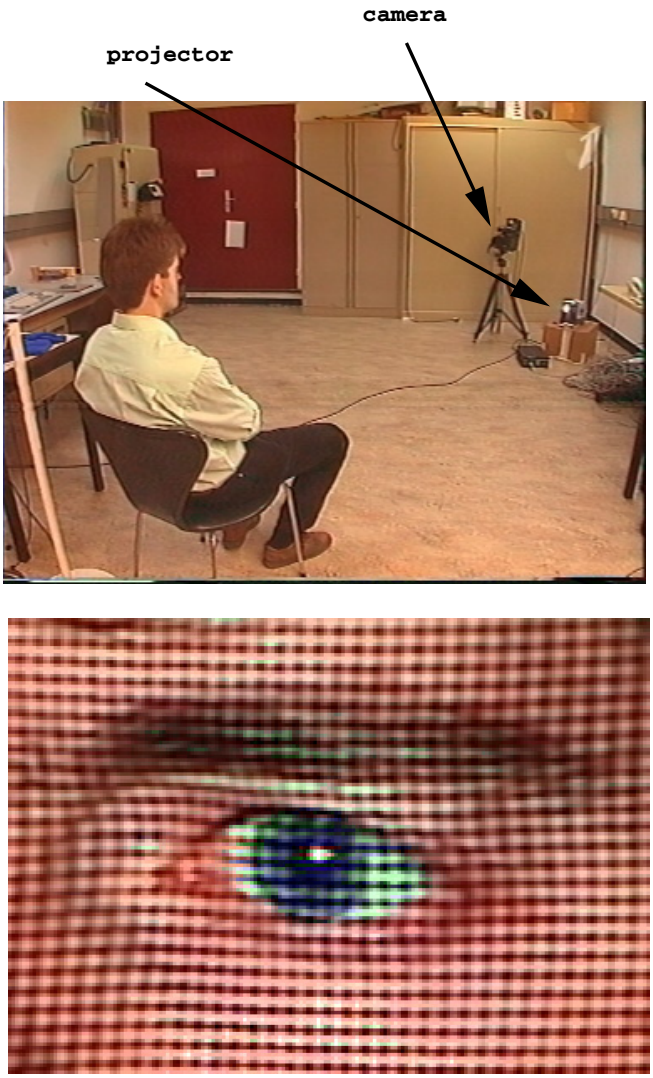

Figure 3. Top: The active system only consists of a normal slide projector and camera, and a computer. Bottom: A regular square pattern is projected on the scene, as seen in this detailed view.

Figure 4. Image with grid for a dwarf figure.

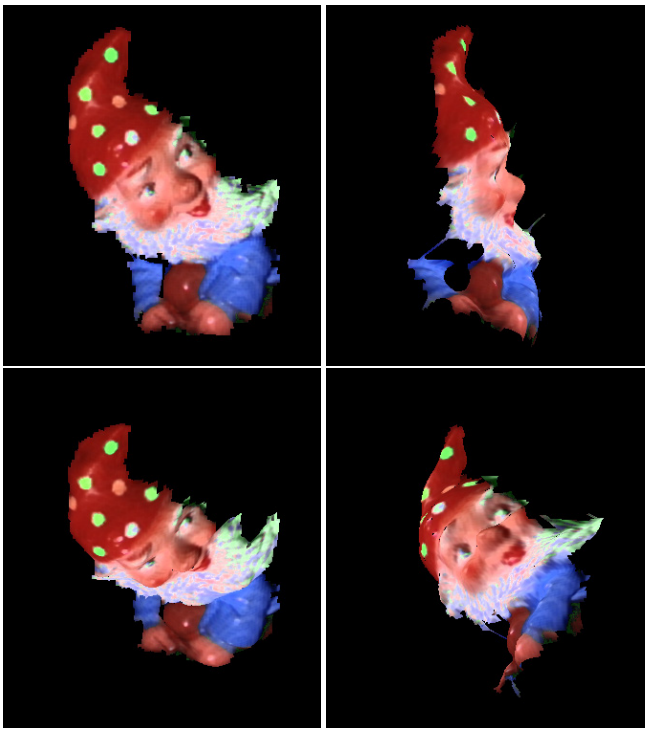

Figure 5. Views of the 3D reconstruction obtained from fig. 4.

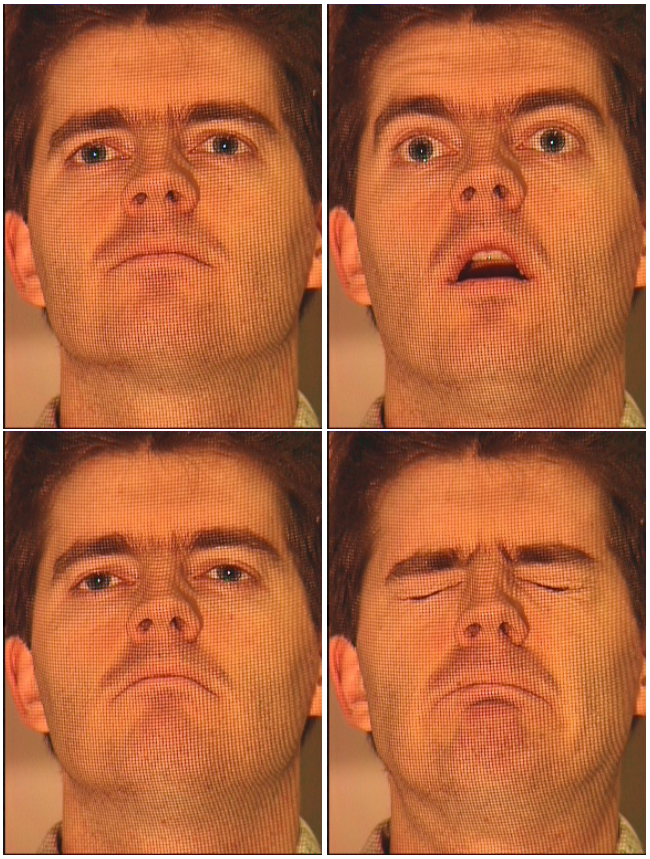

Figure 6. Four frames out of a video with facial expressions.

position is maximised that way.

Fig. 6 illustrates the capacity to capture dynamic 3D. It shows 4 frames of a video of facial expressions. Throughout this sequence the grid was projected. Hence, for every frame a 3D reconstruction could be made. For the moment, the reconstructions for the frames are still carried out independently. This can be improved by the introduction of temporal continuity. Fig. 7 shows reconstructions in a VR setting. All images are composed of the face reconstruction superimposed on a flat background (the background is a simple video taken from an aquarium), and for some frames a virtual air bubble has been integrated. The global head motions, deformations, and changes in shading are all generated from the 3D reconstructions obtained from the initial image sequence. In that sequence the head didn't rotate and the illumination was fixed. The result is a little 3D movie.

Typical computation times for 3D shape are about 2 minutes for a frame on an SGI Indigo, and about 1.5 minutes to add texture if it has to be retrieved from the same image (i.e. with the pattern on). Undoubtedly, these times can be

Figure 7. Example of a dynamic 3D sequence used for a short movie.

reduced further. The code can be optimised and also the exploitation of temporal continuity in time sequences is expected to yield a substantial reduction in computation times. Nevertheless, it is clear that the system in its current form is far from real-time, i.e. cannot perform the reconstructions at video rate. The data can be taken at such a rate, for off-line reconstruction, but processing can't keep up with data acquisition. Currently, the software is being mapped onto hardware as to make it one or two orders of magnitude faster.

4 Conclusions and future work

The basic thesis of this paper has been that new applications of 3D modeling, i.c. the shift from metrology/inspection to visualisation/communications, calls for new 3D acquisition techniques. Two examples were discussed. The first is based on 3D reconstruction from uncalibrated video data. The second uses special illumination but still requires little investment. A normal camera and projector suffice. That system is also very simple to calibrate.

For the future our group plans several extensions of this work. For one thing, it is necessary to further streamline the 3D reconstruction work of larger scenes. Secondly, work is planned on the compression of 3D shapes and the efficient representation of textures. Thirdly, it stands to reason to combine several of such approaches into a single system, that fuses information from the different sources and lets one approach support the other.

Acknowledgements: The authors gratefully acknowledge support of ACTS project AC074 'VANGUARD' and IUAP project 'Imechs', financed by the Belgian OSTC (Services of the Prime Minister, Belgian Federal Services for Scientific, Technical, and Cultural Affairs) Marc Proesmans and Marc Pollefeys thank the IWT (Flemish Institute for the Advancement of Science in Industry) for support through their research grants.

References

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

[2] 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

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

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

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

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