Experimental 3-D digital TV studio

W.P. Cockshott, S. Hoff and J.-C. Nebel

Abstract: The Michelangelo project at the University of Glasgow has developed an experimental three-dimensional television studio. This uses 24 video cameras and parallel computers to capture moving three-dimensional models of human actors. This allows the capture in real time of the appearance and three-dimensional positions of a human actor. It does this using stereo imaging techniques that have been under development at the University of Glasgow for several years. The development of the studio has thrown up many technical problems which are still to be fully resolved, nonetheless it is already producing convincing animated sequences.

1 Introduction

The Michelangelo project at the Universities of Glasgow and Edinburgh was funded by the Scottish Executive and had as its objective the development of advanced three-dimensional (3-D) capture facilities for use by medical science [1], ergonomics and digital media. It involved the development by the University of Glasgow of a 3-D capture system designed for rapid, repeated high-speed 3-D image capture of human bodies: a dynamic scanner.

The basic concept of the dynamic scanner, based on over a decade of research into 3-D imaging by Turing Institute and the Imaging Faraday Partnership, was to equip a studio space such that the 'working volume' is imaged from all directions using fixed stereo-pairs of TV cameras. The stereo-pair images collected by the cameras are then processed using a 3-D imaging software based on photogrammetric techniques developed by the Turing Institute [2, 3], Glasgow, to create a spatiotemporal 3-D model of this space. This would give a full 3-D model of all the action, which can be viewed from any direction. It would also be possible to build a data structure that accommodates information about the objects in 3-D space and their change over time: a true 3-D movie.

2 State of the art

Lead by medical and military researches, the use of 3-D body scanners has begun to become a mature technology and many different techniques are used to generate 3-D static photo-realistic models of real human [4–9]. The main difference between the results these full body scanners provide is about the type of data they can capture. Most commercial scanners, based on laser beams and structured light, have a capture time of about 15 s, whereas the ones using one set of photographs only need milliseconds and can be used to capture moving subjects.

Few research laboratories have developed such scanners. Monks designed a colour-encoded structured light range-finder capable of measuring the shape of time-varying surfaces, where structure light is continuously projected from a single direction [10]. The main application was about measuring the shape of the human mouth during continuous speech sampled at 50 Hz. The Lawrence Livermore National Laboratory has also developed a dynamic scanner using CCD video cameras for medical applications. Their technology dubbed 'CyberSight' is also based on the projection of a specific pattern [11]. Although these two laboratories have successfully generated 3-D sequences, their technologies have some limitations compared to ours because there are based on the projection of a specific known pattern, which cannot be projected on an object from different directions. Therefore, they cannot get a full coverage of 3-D objects.

The work which may be the closest to ours has been developed at the Robotics Institute of Carnegie Mellon University. They are mainly interested in analysing 3-D human motion [12]. For that purpose they built a '3-D room' which is a facility for 4-D digitisation: 49 synchronised video cameras [13] are mounted on the wall and ceiling of the room. 3-D shapes are reconstructed by using a volumetric method called 'shape from silhouette' [14] that creates a voxel model that is then converted to a surface model by using marching cubes.

3 Apparatus

The apparatus developed incorporates many of the design lessons learned from the equipment developed by the Turing Institute [2]. However, the requirements of video capture has meant that new challenges have to be met.

The basic component of the system is termed a 'pod' and comprised a group of three digital cameras, two of which are monochrome and one is colour. The monochrome cameras are used to extract range data and the colour camera is used to record the texture. The entire rig contains eight pods; see the schematic arrangement in Fig. 1.

The algorithm used to extract range information attempts to find the disparity between the positions of a

IEE Proceedings online no. 20030156

DOI: 10.1049/ip-vis:20030156

Paper first received 26th March and in revised form 6th September 2002

The authors are with the Imaging Faraday Partnership, Department of Computing Science, University of Glasgow, 17 Lilybank Gardens, Glasgow G12 8QQ, UK

Techset Composition Ltd, Salisbury
Doc: h:/Journals/Iee/vis/VIS32116.3d

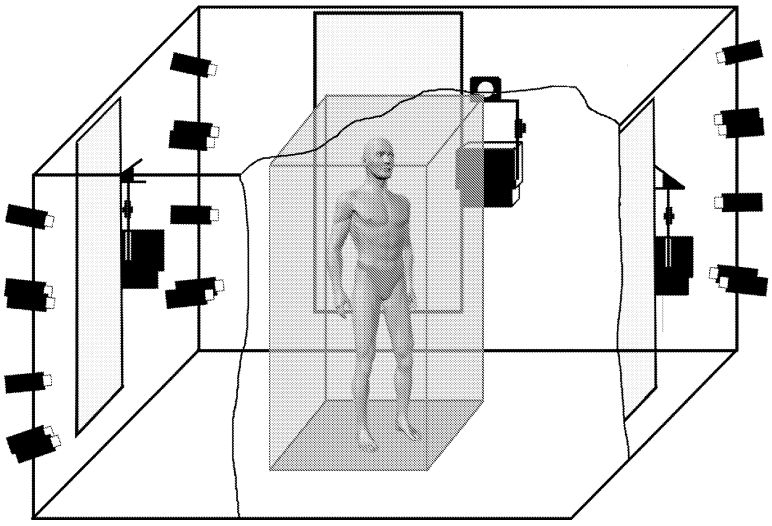

Fig. 1 *Layout of the studio*

given real-world object as viewed from two cameras. Within two given images of a stereo pair, it tries to find the maximally correlated local neighbourhood in these images. For example, as skin is relatively featureless at the pixel spacing and digital precisions available using video cameras, the resulting correlation surfaces can lack well-defined maxima. The scanner overcomes this problem by projecting a pattern image onto the subject.

Earlier scanners used continuous pattern illumination, which was flooded out by flash when texture was captured. That would have required the replacement of the flashes with flood strobes. It proved difficult to obtain strobes sufficiently bright to drown out projected patterns, and so the inverse arrangement was chosen. We use strobe illumination to provide the pattern and conventional studio lighting for the texture pictures. If the pattern was to be visible to the ranging cameras against the ambient light, their exposure times had to be as short as the time of the flashes. The very brief exposure times of about 400 μs, meant the use of very light-sensitive cameras. While colour cameras could have been used throughout, their sensitivity is not as good as that of monochrome cameras, due to the loss of light in their colour filters.

Each pod has one JAI CV-M70 and two Sony XC55 cameras, all with 640×480 pixel resolution. Associated with each pod are strobe lamps fitted within a modified overhead projector and a PC with two frame grabber cards: Coreco Viper RBG and Coreco Viper Quad cards.

4 Optical capture process

The extraction of range information requires the object to be illuminated with a pattern. The best results are achieved by projecting a binary random pattern so that its dots in the image on the imaging chip respect the Nyquist limit. The size of the pattern-slide was estimated by knowing the architecture of the chip, the focal length of camera and projector, the distances from subject to camera and to projector.

As the pattern is projected onto a 3-D object, the quality of the observed pattern depends on the projector's depth of field, hence on the ability to maintain a sharp pattern on parts of the subject that are outside the focal plane. That depth of field depends on several factors, including size of the aperture, size and distance of the light source, and the distance at which the projector should be focused.

By reducing the aperture using diaphragms, we can choose only the ray bundles near to the optical axis. This gives smaller angles on the imaging side of the system, and so a larger depth of field around the focal plane. The disadvantage is that stopping down the aperture by one stop results in halving the amount of light. A lens with a long focal length also helps to increase the depth of field, because it produces smaller angles than one with a short focal length.

Moreover, the best depth of field can be created by a point-like light source. Unfortunately, the strobe tubes-measure $2\,\text{cm} \times 3\,\text{cm}$. If we could place the tube far enough from the projecting lens, it would appear point-like. However, light losses make this option infeasible.

Therefore, a compromise between depth of field and light output has to be found: a stroboscope built into an overhead projector was the easiest most inexpensive possibility. A mirror with an angle of 45° to the optical axis allows a reasonable distance to the first lens, the Fresnel lens of the projector. The projecting lens is replaced with a long focal one.

The active volume in the head set-up is a 50 cm cube. The focus of the projected pattern is ideally set to the first third of the space in which the object is expected to move, because the acceptable depths, of field are from 1/3 before the focus plane to 2/3 behind it.

Each pair of monochrome cameras is mounted one on top of the other 45 cm apart. This distance is the basis of a triangulation-based 3-D capture method: if that distance is too small, small changes in distance to the camera cannot be detected; if is too big, occlusions occur because one camera cannot see the same details as the other. The effects of such occlusions are poor matching forcing interpolations or incomplete models.

The colour camera is mounted next to the monochrome cameras to ensure they have similar views. The incandescent white lights providing the illumination for the colour cameras are set up behind white diffusion screens to reduce highlights. These have to be avoided because the reconstructed 3-D models may, subsequently, be placed in a virtual environment with different lighting conditions.

5 Image acquisition and processing

Software was developed to control the display and the recording of images captured by the 24 video cameras of

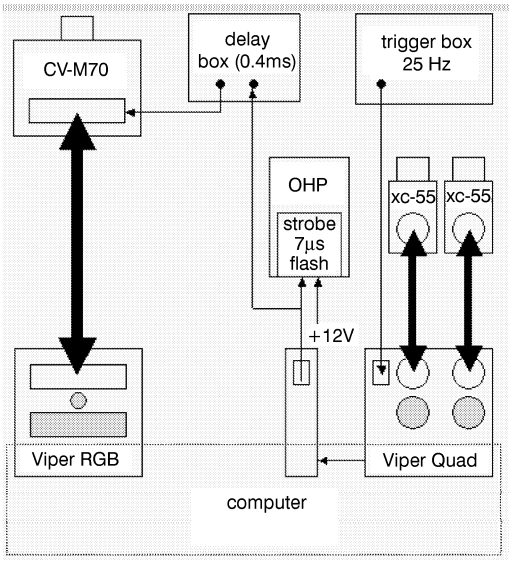

Fig. 2 *Interconnection of pod components*

the system. This software also allows control of the processing of data for calibration or 3-D model generation. As the cameras are connected to eight different computers, communications through network are needed. Moreover, when synchronisation is critical, i.e. for the recording of data, software commands have to be supplemented by a hardware synchronisation device: each pod has a feed from a *master sync* signal operating at 25 Hz, see Fig. 2.

The sequence of stages required to capture image with the system is:

(i) The operator activates a grab on the console of one of the computers.
(ii) The control software tells the frame grabbers for all cameras on all machines currently being used to wait for trigger signals, before capturing a sequence of frames.
(iii) The operator throws a switch on the trigger box to start pulses.
(iv) The trigger is input to the Viper Quad frame grabbers, which trigger the monochrome cameras and the strobe projectors. The shutters are open for 400 μs.
(v) The strobe trigger is also passed through a Thurlby Thandar TGP 110 digital pulse generator operating in triggered delay mode. This allows a controllable delay to be inserted into the signal. We currently use a delay of 400 μs.
(vi) The output from the TGP110 is sent to all of the colour cameras to trigger their shutters to open. These open for 4 ms.
(vii) Images are then downloaded from all cameras to the frame grabbers before the next *master sync* pulse.
(viii) The frame grabbers perform a direct memory access transfer of the images over the PCI bus into the system area of main memory on the PC.

The process of 3-D data capture involves very large data flows. Data arrives from the frame grabbers at a combined data rate of 307 MByte/s. A twenty-second 'take' thus generates over 6 GByte of data. This data rate is in excess of the sustained throughput of the disks attached to the processors. It is thus necessary to buffer the data in RAM as it is being captured. For sequences longer than 2.5 s, data have to be transferred from system to user RAM by the CPU. This configuration is illustrated in Fig. 3.

Once the data have been saved, that network of computers is used as a parallel machine to process the data: 3-D data extraction and mesh generation tasks are dispatched over the network.

6 3-D model generation

The process of extracting 3-D information from a stereo-pair of images is termed 'stereo matching'. The matching algorithm used was developed by Jin, and is based on multiresolution image correlation [15, 16].

The algorithm takes as input a pair of monochrome images and outputs a pair of images specifying the horizontal and the vertical displacements of each pixel of the left image compared to the matched point in the right image. The matcher is implemented using a difference of Gaussian image pyramid: the top layer of the pyramid is 16 by 12 pixels in size for a base of 640 by 480 pixels. Starting from the top of the pyramid, the matching between the two pictures is computed. Then using the displacements, the right image of the next layer of the pyramid is warped to fit the left image. Thus, if the estimated disparities from matching at the previous layer were correct, the two images would now be identical, occlusions permitting. To the extent that the estimated disparities were incorrect there will remain disparities that can be corrected at the next step of the algorithm, using information from the next higher waveband in the images.

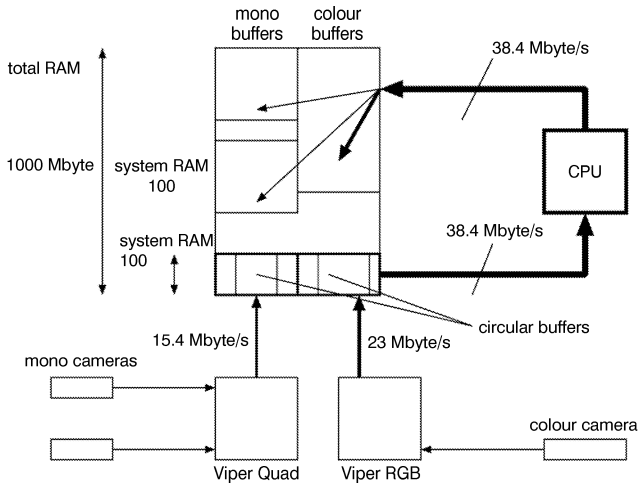

Fig. 3 *Dataflow during combined monochrome and colour capture*

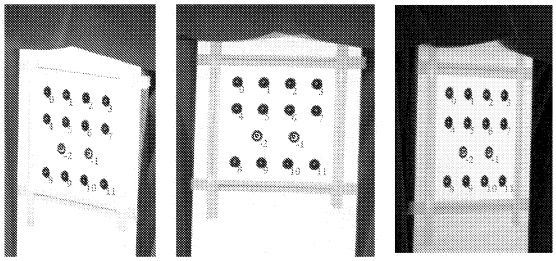

Fig. 4 *Calibration target with the circles detected by the software*

As, at each layer, the two images are supposed to match more or less, thanks to the warping step, correlation on 5 by 5 pixel neighbourhoods is sufficient to find pairs of corresponding pixels. Once the matching process is completed, the final displacement files combined with the calibration file of the stereo system allow the generation of a range map.

Once 3-D data have been extracted for each pod, these data need to be merged to generate a 3-D model. As the pods have been calibrated together, the eight sets of 3-D data captured at a given time step can be integrated in a single co-ordinate frame. An implicit surface is then computed that merges together the point clouds into a single triangulated polygon mesh, using a variant of the marching cubes algorithm [17]. The generation of photo-realistic models is achieved by mapping the colour pictures taken by the colour cameras to the 3-D geometry. As these cameras were calibrated in the same co-ordinate frame as their corresponding black and white pairs, each 3-D point can be associated to a value on the colour image. However, as the colour balance and the settings of the eight cameras are not identical, it is necessary to process the images to get a seamless texture for the 3-D mesh. First, the images are processed globally to get similar colour balances, then a local processing is applied for pixels, which are in contact on the 3-D models, but belong to different images. Finally, the output of our system is a sequence of VRML files with their associated JPEG textures files.

7 Calibration

As we use nonmetric cameras, the system has to be calibrated before utilisation to extract the absolute intrinsic and extrinsic parameters of every camera that is used. To do so, a calibration plain target, Fig. 4, consisting of twelve black circles and two black rings of accurately known dimensions and locations, is presented to all the cameras; when the target is seen by several cameras, the position and orientation of these cameras can be determined with respect to each other. Software then locates the contour, centre and location on the target for every circle. This computation allows the generation of an approximate geometric model of each camera and its orientation to the target by using the direct linear transform (DLT) [18]. From this model, a much more accurate model is computed. It is termed space resection by photogrammeterists and the general algorithmic approach is contained within a process known as bundle adjustment [19].

Calibration experiments indicate the ranging accuracy to be 1/160 of the distance from the pod to the subject [16].

8 Sample results

The sample images are of the actor Jeremy Killick playing the part of Napoleon in an experimental film. Fig. 5 shows the process by which a set of three captured images from a pod are converted into distance information. The images labelled top and bottom come from the two monochrome cameras. The random pattern can be observed. The images labelled horizontal and vertical disparity map encode the distances in the horizontal and vertical directions between corresponding pixels in the top and bottom images. The confidence map image records the correlation coefficients over 5 by 5 pixel windows, centred on pixels in the bottom image, and the corresponding pixel position in the top image. In conjunction with camera calibration data, a depth image is produced. Fig. 6 shows several different views of Napoleon's head all taken from the same 3-D model. The centre image had the skin texture partly removed to show the underlying model. The right-most image shows the triangulated mesh of the model. Rotation is used to illustrate the 3-D nature of the model. Fig. 7 shows five models selected from one take of 625 frames. Fig. 8 is taken from a sequence of models showing a jump. The artefacts around the body are due to the camera resolution, their distance to the subject and the background in the images.

9 Conclusion

In this paper we have given a detailed description of the dynamic 3-D whole body scanner we are currently developing, where original software and hardware systems have been investigated. We have demonstrated the validity of the concept by generating a true 3-D film, using a configuration for head scanning.

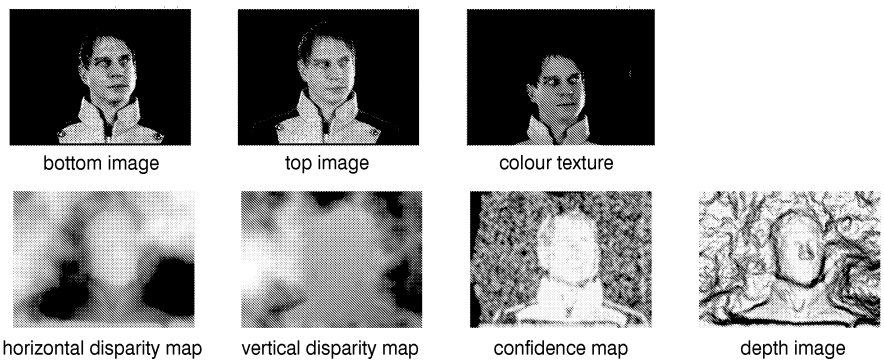

Fig. 5 *Derivation of individual range data from images*

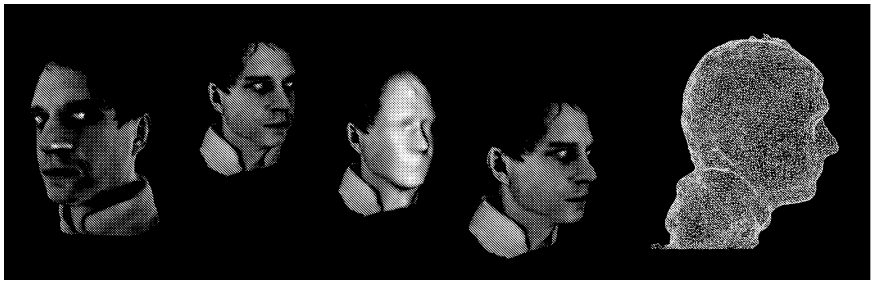

Fig. 6 *Different views of an individual model*

Fig. 7 *Sequence of models captured over time*

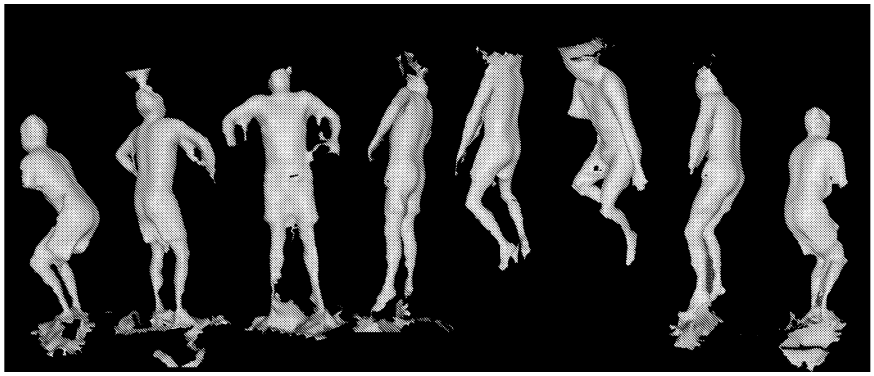

Fig. 8 *Sequence of models without texture over time*

In the future, we will address the two main limitations of our system: the illumination of the subject and the data flow generated.

First, we plan to change the illumination to a continuous pattern illumination based on a shadow throw to overcome the depth of field problems. Moreover, we plan to use powerful stroboscopes to over flash the pattern illumination and to provide better colour images. The strobes should flash with 100 Hz, to ensure a better comfort to the scanned subject.

Secondly, we intend to tackle our data flow issues by researching better transfer protocols and increasing the size of our computing power using a bigger network of computers.

10 Acknowledgments

This work was supported by the SHEFC project 'Michelangelo'. The authors would like to thank Kim Bour (director), Philip Lloyd (producer), Andrew Bampfield (writer) and Madeleine Bowyer and Jeremy Killick (actors) for giving them the opportunity of scanning the first 3-D film.

11 References

1 AYOUB, A.F., SIEBERT, J.P., WRAY, D., and MOOS, K.F.: 'A vision-based three dimensional capture system for maxillofacial assessment and surgical planning', *Brit. J. Oral Maxillofacial Surg.*, 1998, **36**, pp. 353–357
2 SIEBERT, J.P., and URQUHART, C.W.: 'C3D: a novel vision-based 3-D data acquisition system'. Proceedings of Mona Lisa European Workshop, Combined Real and Synthetic Image Processing for Broadcast and Video Production, 1994, Hamburg, Germany
3 URQUHART, C.W: 'The active stereo probe, the design and implementation of an active videometrics system'. PhD Dissertation, Turing Institute and the University of Glasgow, 1997
4 SIEBERT, J.P., and MARSHALL, S.J.: 'Human body 3D imaging by speckle texture projection photogrammetry', *Sens. Rev.*, 2000, **20**, (3), pp. 218–226
5 TRIEB, R.: '3D-Body Scanning for mass customized products - Solutions and Applications'. Proceedings of International conference of numerisation 3D-Scanning, 2000
6 CYBERWARE, http://www.cyberware.com
7 WINSBOROUGH, S.: 'An insight into the design, manufacture and practical use of a 3D-body scanning system'. Proceedigs of International conference of numerisation 3D-Scanning, 2000

8 VAREILLE, G.: 'Full body 3D digitizer'. Proceedings of International conference of numerisation 3D-Scanning, 2000
9 TCTi, http://www.tcti.com
10 MONKS, T.P.: 'Measuring the shape of time-varying objects'. PhD Dissertation, University of Southampton, 1994
11 Lawrence Livermore National Laboratory, http://www.llnl.gov/automation-robotics/cyber.html
12 KANADE, T., VEDULA, S., BAKER, S., RANDER, P., and COLLINS, R.: 'Three-dimensional scene flow'. Proceedings of the 7th International Conference on Computer Vision, 1999
13 KANADE, T., SAITO, H., and VEDULA, S.: 'The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams'. Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 1998
14 VEDULA, S., RANDER, P., SAITO, H., and KANADE, T.: 'Modeling, combining, and rendering dynamic real-world events from image sequences'. Proceedings of Fourth International Conference on Virtual systems and multimedia, Gifu, Japan, 1998
15 ZHENGPING, 'On the multiscale iconic representation for low-level computer vision systems'. PhD dissertation, Turing Institute and the University of Strathclyde, 1998,
16 NEBEL, J.C., RODRIGUEZ-MIGUEL, F.J., and COCKSHOTT, W.P.: 'Stroboscopic stereo rangefinder'. Proceedings of 3DIM2001, Québec City, Canada, 2001
17 LORENSEN, W.E., and CLINE, H.E.: 'Marching cubes: a high resolution 3D surface construction algorithm', *Comput. Graph.*, 1987, **21**
18 ABDEL-AZIZ, Y.F., and KARARA, N.M.: 'Direct linear transformation from comparator coordinates into object coordinates in close-range photogrammetry'. Proceedings of ASP Symposium on close-range photogrammetry, Illinois, January 1971, pp. 1–18
19 KARARA, H.M.: 'Handbook of non-topographic photogrammetry' (American Society for Photogrammetry and Remote Sensing, Falls Church, 1989, 2nd edn.)