An Evaluation of Three Popular Computer Vision Approaches for 3-D Face Synthesis

Alexander Woodward, Da An, Yizhe Lin, Patrice Delmas,
Georgy Gimel'farb, and John Morris

Dept. of Computer Science, Tamaki Campus
The University of Auckland, Auckland, New Zealand
awoo016@ec.auckland.ac.nz,
{p.delmas, g.gimelfarb, j.morris}@auckland.ac.nz

Abstract. We have evaluated three computer approaches to 3-D reconstruction - passive computational binocular stereo and active structured lighting and photometric stereo - in regard to human face reconstruction for modelling virtual humans. An integrated experimental environment simultaneously acquired images for 3-D reconstruction and data from a 3-D scanner which provided an accurate ground truth. Our goal was to determine whether today's computer vision approaches are accurate and fast enough for practical 3-D facial reconstruction applications. We showed that the combination of structured lighting with symmetric dynamic programming stereo has good prospects with reasonable processing time and accuracy.

1 Introduction

Vision based 3-D facial reconstruction is appealing because it uses low-cost off-the-shelf hardware. Our main objective was to assess the usability of three of the most popular reconstruction techniques - computational binocular stereo, structured lighting and photometric stereo - for creating realistic virtual humans. Binocular stereo is of particular interest as it is a passive technique, whereas the other two actively project light onto the scene. Determining whether a passive approach can provide results competitive with active techniques is important.

Seeing and interacting with humans is commonplace in a person's everyday life. Indeed, most verbal and non-verbal communication uses part of the face. Facial modelling has therefore become a major issue for the successful design of human computer interfaces. The applications for facial modelling to create virtual humans are wide and varied, including surveillance, entertainment and medical visualisation [10]. Faces are highly emotive and consequently virtual humans are a powerful tool, often a necessary one, in a variety of multimedia applications.

Section 2 briefly surveys the state-of-the-art in face reconstruction techniques. Accuracy criteria relevant to face reconstruction and vision based 3-D reconstruction techniques are summarised in Section 3. The experimental setup is described in Section 4, Sections 5 and 6 discuss experimental results.

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 270–278, 2006.

2 Previous Work

Facial reconstruction is a very specific task. Image based 3-D reconstructions appear most accurate when viewed under directions similar to those in which they were acquired. Rotations to novel views of the 3-D data often reveal the most prominent flaws in a reconstruction. However, performance analysis of vision based reconstruction has focused on a collection of arbitrarily chosen scenes [9]. We focused on human face reconstruction because of its identified importance. The techniques compared here have been described in detail [3,4,7,11,12].

Facial reconstruction from digital images reduces modelling time and allows for a personalised result. Almost all vision based techniques use a generic face model that is warped to the raw data.

Successful techniques [8] use data gathered from a 3-D scanner. Unfortunately the cost of 3-D scanning equipment makes this impractical for many situations.

3 Tested Reconstruction Algorithms

In contrast to previous work, we focus on more stringent error analysis and criteria for face reconstruction. The characteristic face feature areas - eyes, mouth, nose, etc - are especially important for reconstruction.

Accuracy of surface normal reconstruction, which is often neglected in existing analysis, is an important indicator of quality when a surface area exhibits an overall shift in depth but retains a low comparative depth variance measure. We included this measure to provide an extended reconstruction error analysis.

There are a large number of algorithms for 3-D reconstruction so we selected some of the most popular techniques in each of the chosen approaches.

Binocular Stereo. After comparing a set of implemented dense two-frame stereo algorithms, we chose the algorithms in Table 1 as they provide a cross-section of local and global techniques. Global algorithms incorporate an optimisation process over the entire domain and produce smoother results, but usually at the sacrifice of speed. Tthe algorithms used are described elsewhere [7,9].

Table 1. Tested Binocular Stereo Techniques

'Winner Takes All' Sum of Absolute Differences (SAD)[1]	- local algorithm
Dynamic Programming Method (DPM)[1]	- global algorithm
Symmetric Dynamic Programming Stereo (SDPS)[2]	- global algorithm
BVZ (Graph Cut based algorithm)[1]	- global algorithm
Belief–Propagation (BP)[3]	- global algorithm
Chen and Medioni (CM)[2]	- local algorithm

[1] Scharstein and Szeliski, http://cat.middlebury.edu/stereo/code.html
[2] Our own implementation
[3] Felzenszwalb and Huttenlocher, http://people.cs.uchicago.edu/∼pff/bp/ [14]

Structured Lighting. Structured lighting techniques use active illumination to label visible 3-D surface points. Active illumination aims to simplify the surface reconstruction problem. Reconstruction time depends on a compromise between the number of images required (for complex coding strategies) and uniqueness of pixel label and thus ability to resolve ambiguities. The Gray code algorithm matches codes whereas both of the direct coding techniques project a light pattern that aids the correspondence process in a standard binocular stereo algorithm, cf. Table 2.

Table 2. Structured lighting techniques to test

Time-multiplexed structured lighting using Gray code
Direct Coding with a Colour Gradation Pattern
Direct Coding with a Colour Strip Pattern

We aim to determine whether a simpler single light projection coupled with a traditional stereo algorithm is competitive with a more complex coding scheme such as a Gray code constructed from multiple projections. An et al. give a more detailed description of the structured lighting techniques used [4].

Photometric Stereo. An albedo independent approach [5] with three light sources was used in this experiment. This technique assumes Lambertian scatterers, a parallel projection model and light sources situated at infinity. However this is a drastic simplification of reality. This paper focusses on assessing the gradient field integration component of photometric stereo. The algorithms were chosen to present both local and global techniques. Global algorithms incorporate an optimisation process over the entire field and produce smoother results. The presented gradient field integration techniques are described by Woodward and Delmas [12].

Table 3. Tested photometric stereo techniques

Frankot-Chellappa Variant (FCV)	- global algorithm
Four-Scan Method	- local algorithm
Shapelets (9 scales)	- local algorithm

4 Experimental Setup

A diagram of each sub-system is in Fig. 1. Images were taken automatically through specifically designed software and all data was processed in a batch manner. For each test subject, the facial region (about 800×700 pixels) was cut from the images for comparison.

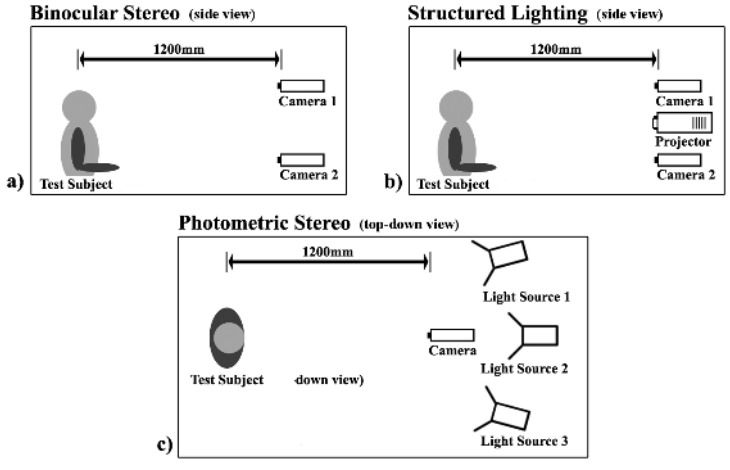

Fig. 1. System geometries for all techniques

A *Solutionix Rexcan 400* 3-D scanner (depth accuracy ∼0.5 mm, planar resolution 0.23 mm) was used to obtain ground truth data for each test subject (see Figure 2).

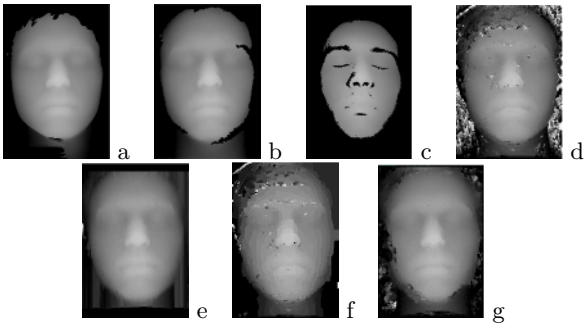

Fig. 2. Reconstruction examples: a) Ground truth, b) Gray code, c) FCV, d) SAD, e) SDPS, f) BVZ, g) CM

4.1 Binocular Stereo

A pair of *Canon EOS 10D* 6.3 Mpixel cameras was used for high resolution image acquisition. This allows for very dense disparity maps and accordingly a larger disparity range. Each camera lens has a measured focal length of 52 mm. The baseline separation between the two cameras was 175 mm. The cameras were aligned with their optical axes parallel, allowing for simplified reconstruction formulae. The test subject was placed approximately 1200 mm from the cameras.

4.2 Structured Lighting

This system used the same cameras as in binocular stereo. The main concern is the slow acquistion time that belies a potentially fast process when the appropriate hardware is available. With these cameras, it is in the order of tens of seconds.

An *Acer LCD Projector, model PL111*, was used to project light patterns into the scene. The device is capable of projecting an image of 800×600 pixels and has a focal length of $21.5 - 28$ mm.

4.3 Photometric Stereo

A system with three $150W$ light sources was used [5]. A *JVC KY-F55B* camera controlled automatically by a switching device connected to a computer captured the images. As shown in Figure 1c, the lights are positioned so as to be non-coplanar which is a requirement for the algorithm to work correctly.

4.4 System Calibration

A cubic calibration object with 63 circular calibration markings distributed evenly over two of its sides was used. Tsai's calibration technique was used [13].

A light calibration step must also be performed for the photometric stereo system. This determines the direction to the lights from an arbitrary scene origin. A calibration sphere was used for this process as directions can be determined analytically. The sphere was placed in the same location as the subject will be positioned during data acquisition.

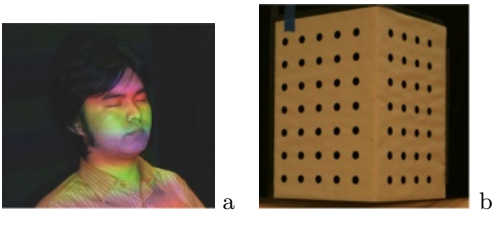

Fig. 3. (a) Test subject during acquisition with a projected colour pattern. (b)Calibration object for camera calibration.

4.5 Image Rectification

Stereo images were rectified by converting them to a standard epipolar stereo geometry. The rectification process transforms and resamples the images so that the resultant image pairs will meet the epipolar constraint.

To satisfy this requirement, an intuitive rectification is to rotate both cameras around their optical centres to a common orientation. The rotated cameras still comply with the pinhole camera model and the baseline remains intact. Using

a calibration result (see Section 4.4), one can compute the baseline and a new common orientation from the pose of the two cameras. This method is similar to the method of Ayache and Hansen [1] which insists on neither the extrinsic nor the intrinsic parameters of a camera but the 3×4 perspective projection matrix. Our method utilises the extrinsic and intrinsic parameters, which is simpler and decouples the lens distortion from the aforementioned 3×4 matrix.

4.6 Data Processing

Data from the several experiments was aligned using a semi-automatic process involving 3-D object rigid transformations using the 3-D scanner software which allows for data manipulation and registration. After alignment all data was subsequently projected into disparity space and disparity maps were compared. Thus our primary accuracy metric was disparity (depth) deviations from the ground truth data. Throughout the experiment, it was found that 3-D data alignment is a difficult process and much care is needed. A small number of correspondences were entered manually to ensure correct registration.

5 Experimental Results

A Pentium 4 3.4 GHz machine with 2 Gbyte RAM computed the depth maps. The resultant face reconstructions and a ground truth of the test subject were compared. A set of 17 subjects were used.

The reconstruction accuracy metrics were: the percentage of pixels with absolute depth errors less than two disparity units ($P_{<2}$), the maximum (max) absolute pixel depth error , the mean (e_{mn}) absolute pixel depth error, the standard deviation (σ_e) of errors, and the mean cosine error (MCE). Central differencing was used to estimate surface normals, and the MCE measures the quality of reconstruction of surface normals:

$$MCE = \left| \left(\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{n}_{i,j} \bullet \mathbf{n}^*_{i,j} \right) - 1 \right| \tag{1}$$

where M, N are the image dimensions, $\mathbf{n}_{i,j}$ and $\mathbf{n}^*_{i,j}$ are the reconstructed surface and ground truth normals, respectively, and "$\bullet$" is the dot product operator. The MCE measures how close the reconstructed surface normals are to the ground truth, in particular, MCE = 0 if $\mathbf{n}_{i,j} = \mathbf{n}^*_{i,j}$, 1 if $\mathbf{n}_{i,j} \perp \mathbf{n}^*_{i,j}$, and 2 if $\mathbf{n}_{i,j}$ and $\mathbf{n}^*_{i,j}$ are collinear but with opposite directions.

The experimental results in Table 4 show that active reconstruction techniques consistently perform better than purely passive ones. Passive binocular stereo is greatly improved by supplementing the process with only a single light pattern (indicated as *Gradation* and *Strip* in Table 4).

Photometric stereo, although active in nature, is unable to recover true depth measurements due to the required gradient field integration step. None of the

Table 4. Average reconstruction accuracy and running time

Method	$P_{<2}$, %	max	e_{mn}	σ_e	MCE	Time, *sec*
Gray code	97	8	0.6	0.6	0.01	4.0
SDPS	89	13	1.0	0.9	0.09	6.0
SDPS + Gradation	90	13	1.0	1.0	0.11	.
SDPS + Strip	93	9	0.8	0.7	0.09	.
DPM	79	19	1.4	1.6	0.24	6.0
DPM + Gradation	84	13	1.2	1.2	0.25	.
DPM + Strip	92	13	0.8	0.8	0.14	.
BVZ	77	42	1.8	3.4	0.12	3517
BVZ + Gradation	83	31	1.3	1.5	0.09	.
BVZ + Strip	92	40	0.9	1.6	0.09	.
SAD	80	42	1.8	3.4	0.17	1.7
SAD + Gradation	85	32	1.2	1.7	0.16	.
SAD + Strip	93	35	0.8	1.3	0.09	.
BP	73	27	2.1	3.0	0.18	180
BP + Gradation	77	21	1.8	2.3	0.16	.
BP + Strip	89	18	1.0	1.2	0.16	.
CM	88	20	1.0	1.1	0.09	30.0
CM + Gradation	89	22	1.2	1.4	0.13	.
CM + Strip	92	21	0.9	1.1	0.10	.
PSM FCV	69	14	1.7	1.7	0.09	4.0
PSM Four-path	54	13	2.4	2.0	0.05	37.0
PSM Shapelet	71	12	1.7	1.7	0.04	153

Gradation and *Strip* refer to active projection of a Colour Gradation or Colour Strip pattern, respectively, on the object.

compared photometric stereo algorithms performed as well as the best offerings found in the other two approaches.

The performance of a pure Gray code approach is clearly superior to other techniques. It attains the lowest scores for all categories. Through effective formulation, it can handle coding errors that can happen in problem areas having low albedo or strong specularities, such as the eye regions [4] where PSM techniques usually fail.

The tuning of parameters is a difficult task. They are usually set with respect to the image size. It was found that global algorithms based on more complex optimisation techniques such as Belief Propagation (BP) [14] and the Graph Minimum Cut (BVZ) [2] did not perform as well as expected for human faces and relatively large disparity ranges. Thus our results differ from Scharstein and Szeliski's ranking of stereo algorithms [9] and the Middlebury Stereo Vision web page (www.middlebury.edu/stereo/). Our test has much higher resolution images and, in turn, much greater depth ranges. On facial images the accuracy of dynamic programming based algorithms was similar or even better than for

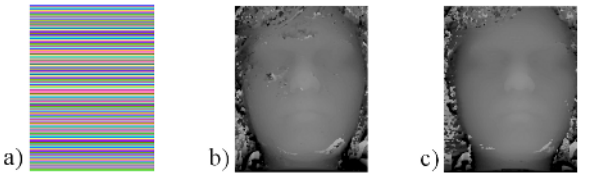

Fig. 4. Stereo (SAD) with and without a projected pattern. a) The colour strip pattern used, b) SAD without projected pattern, c) SAD with projected pattern.

these much more computationally complex (and supposedly better performing) BP and BVZ algorithms.

Colour projections that are similar to skin tones should be avoided in order to provide maximal contrast over the facial surface. The spatial frequency of projected patterns is important and needs to be high enough to provide uniqueness in matching. Thus a low frequency gradation pattern does not perform as well as a strip pattern.

6 Conclusion and Future Work

We introduced a framework and test bench for passive and active 3-D acquisition systems using three different approaches (binocular stereo, photometric stereo and structured lighting) and sixteen algorithms. We compared the data acquired to a benchmark with sub milli-metre depth accuracy using surface normal and depth map information.

All tested algorithms showed reconstruction errors that exceed the requirement for direct presentation of virtual humans and this is currently only remedied in postprocessing steps. Our experiments have shown that errors do not occur in specific areas of the face. Masking out specific regions that are highly textured, counter lowly textured, does not cause significant alterations in results.

Active methods such as structured lighting and photometric stereo have problems with specular, shadow and low albedo regions. Binocular stereo has problems dealing with texture-less regions of the face, the projection of a colour strip pattern saw a marked improvement in reconstruction accuracy. This can be easily seen in the example presented in Figure 4. The FCV algorithm performs at the forefront of the tested PSM algorithms when considering both accuracy and time complexity. Overall, the Gray code approach provides the expected best overall results. However, from these results it appears that the SDPS algorithm coupled with just a single strip pattern is a strong choice in terms of accuracy and time complexity.

We are currently assessing further algorithms, especially those for binocular stereo. The combination of active illumination and stereo vision (using the SDPS algorithms) shows the best potential for generating 3-D characters from a rig of video-cameras.

References

1. N. Ayache and C. Hansen. Rectification of Images for Binocular and Trinocular Stereovision. In *Proc. 9th Int. Conf. on Pattern Recognition, Rome, 1988*. IEEE CS Press: Los Alamitos, 1988.
2. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23(11), pp. 1222–1239, 2001.
3. M. Chan, P. Delmas, G. Gimel'farb, and P. Leclercq. Comparative study of 3d face acquisition techniques. In A. Gagalowicz and W. Philips, eds., In *Proc. Int. Conf. Computer Analysis of Images and Patterns (CAIP'05), Versaille, France*, LNCS 3691, pp. 740–747, Sept. 2005.
4. D. An, A. Woodward, P. Delmas, and C. Chen. Comparison of Structured Lighting Techniques with a View for Facial Reconstruction. In *Proc. Image and Vision Computing New Zealand Conf.* , Dunedin, New Zealand, pp. 195–200, 2005.
5. R. Klette and K. Schluns. *Computer Vision - Three-dimensional Data from Images*. Springer: Berlin, 1998.
6. T. Kurihara and K. Arai. A transformation method for modeling and animation of the human face from photographs. In *Proc. Computer Animation'91 Conf., Tokyo*, pp.45–58, 1991.
7. P. Leclercq, J. Liu, M. Chan, A. Woodward, G. Gimel'farb, and P. Delmas. Comparative study of stereo algorithms for 3D face reconstruction. In *Proc. Int. Conf. on Advanced Concepts for Intelligent Vision Systems, Brussels, Belgium*, Sept. 2004.
8. Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proc. ACM SIGGRAPH'95 Conf.*, pp.55–62, 1995.
9. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, vol. 47(1), pp. 7–42, 2002.
10. Z. Wen and T.S. Huang. 3D Face Processing: Modeling, Analysis and Synthesis. *The International Series in Video Computing*, Vol. 8, Springer: Berlin, 2004.
11. A. Woodward and P. Delmas. Towards a low cost realistic human face modelling and animation framework. In *Proc. Image and Vision Computing New Zealand, Akaroa, Christchurch, New Zealand*, pp. 11–16, Nov. 2004.
12. A. M. Woodward and P. Delmas. Synthetic Ground Truth for Comparison of Gradient Field Integration Methods for Human Faces. In *Proc. Image and Vision Computing New Zealand, Dunedin, New Zealand*, pp. 155–160, Nov. 2005.
13. R.Y. Tsai. A Versatile Camera Calibration Technique for High Accuracy 3-D Machine Vision Metrology using Off the Shelf TV Cameras and Lenses. *Int. J. Robotics and Automation*, vol. 3(4), pp. 323–344, 1987.
14. P.F. Felzenszwalb, and D.P. Huttenlocher. Efficient Belief Propagation for Early Vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, IEEE CS Press: Los Alamitos, pp. 261–268, Jun. 2004.