# A Comparison of Three 3-D Facial Reconstruction Approaches

Alexander Woodward, Da An, Georgy Gimel'farb, Patrice Delmas
Dept. of Computer Science, Tamaki Campus, The University of Auckland
Auckland, New Zealand
awoo016@ec.auckland.ac.nz

## Abstract

*We compare three Computer Vision approaches to 3-D reconstruction, namely passive Binocular Stereo and active Structured Lighting and Photometric Stereo, in application to human face reconstruction for modelling virtual humans. An integrated lab environment was set up to simultaneously acquire images for 3-D reconstruction and corresponding data from a 3-D scanner. This allowed us to quantitatively compare reconstruction results to accurate ground truth. Our goal was to determine whether any current Computer Vision approach is accurate enough for practically useful 3-D facial surface reconstruction. Comparative experiments show the combination of Structured Lighting with Symmetric Dynamic Programming based Binocular Stereo has good prospects due to reasonable processing time and sufficient accuracy.*

## 1. Introduction

Seeing and interacting with faces is commonplace in a person's everyday life. The areas of application for facial modelling to create virtual humans are wide and varied, appearing in such areas as security, the entertainment industry, and medical visualisation. Faces are highly emotive and consequently virtual humans are a powerful tool, often a necessary one in a variety of multimedia applications.

Vision based 3-D facial reconstruction approaches are appealing due to their general low-cost usage of off the shelf hardware. Three of the most popular approaches chosen for study are Binocular Stereo, Structured Lighting, and Photometric Stereo. Their usability for creating realistic virtual humans is the main objective of this work. The performance of Binocular Stereo is of particular interest as it is a passive technique, whereas the other two actively project light onto the scene.

Due to the familiarity of faces we are very sensitive to any nuances or oddity that could appear in a virtual representation. This specificity presents a unique challenge for 3-D reconstruction. The form of the face can generally be defined as a near-homogeneous curvilinear surface. Passive image based techniques must deal with large regions of lowly textured skin regions. Subsequently, knowledge of facial properties can improve reconstruction accuracy.

The subsurface scattering of light within the skin and the anisotropic reflection and specularity of hair proves a difficultly for Photometric Stereo. Such photo-inconsistencies manifest as noise in the correspondence process of Binocular Stereo, or code errors in Structured Lighting.

Section 2 below overviews the current state-of-the-art in facial reconstruction techniques. Accuracy criteria relevant to face reconstruction and vision based 3-D reconstruction techniques are summarised in Section 3. The lab setup is given in Section 4, and Sections 5 and 6 discuss results.

## 2. Previous Work

Facial reconstruction is a very specific task. Image based 3-D reconstructions appear most accurate when viewed under directions similar to that in which they were acquired. Rotations to novel views of the 3-D data often reveal the most prominent flaws in a reconstruction. However, the majority of analysis on vision based reconstruction has focused on general performance for arbitrary scenes [13].

Computer vision based facial reconstruction reduces modelling time and allows for a personalised result. There exist many successful yet manually involved reconstruction techniques, e.g. [7, 12]. Currently automatic techniques that are both robust and accurate are still in their infancy. Almost all vision based techniques leverage the use of a generic face model that is warped to the raw data. The primary difference involves what information is obtained from any input image(s) and how this is used to alter the generic model. Often these methods leverage statistical or heuristic knowledge of faces to aid in their operation.

Successful techniques such as those in [10] and [16] use data obtained from a 3-D scanner. Unfortunately the price of 3-D scanning equipment makes this impractical for most lab situations.

Among the Computer Vision techniques, purely feature based ones have been pursued that obtain a sparse data set. 3-D information is then inferred through considering 3-D edge features [4], feature assignments in video streams [11] and orthogonal views [9]. Binocular stereo producing a dense 3-D data set has been applied for faces in [3, 14].

3-D databases of heads have been used to understand the statistical variance of facial proportions. A 3-D model can then be obtained from features of a single image [5].

## 3. Reconstruction Algorithms for Testing

In comparing to known work, we focus on more stringent error analysis and criteria for face reconstruction. The characteristic face feature areas such as the eyes, mouth, nose, are especially important for reconstruction. These areas carry most of the audio-visual information expressed by humans and their known locality should be considered in any technical evaluation. However, previously conducted anthropometric quantitative analysis has been inconclusive [1].

The accuracy in surface normal reconstruction, one 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 include this measure to provide an extended reconstruction error analysis.

Due to a rich variety of Computer Vision based algorithms for 3-D reconstruction, we test a few most popular techniques in each of the chosen domains.

**Binocular Stereo** After comparing a set of implemented dense two-frame stereo algorithms, we have chosen the algorithms in Table 1 to 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 in execution time.

The trade-offs between accuracy and time complexity are of importance when dealing with large images. For example, top-performing graph cut based techniques have notably higher time complexity, and this is an important issue when dealing with the issue of practicality for the size of images dealt with in this study (see Sections 4 and 5).

This set is currently being expanded upon to incorporate more of the latest Binocular Stereo algorithms. For a review of the presented algorithms see [8, 13].

**Structured Lighting** Structured Lighting techniques use active illumination to code visible 3D surface points. Reconstruction time depends on a compromise between the number of images required (for the case of complex coding strategies) and subsequent uniqueness in pixel code. From Table 2, 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.

**Table 1. Tested Binocular Stereo Techniques**

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

[1] from http://cat.middlebury.edu/stereo/code.html

[2] our own implementation

[3] from http://people.cs.uchicago.edu/∼pff/bp/, [17]

**Table 2. Structured Lighting Methods to Test**

| |
|---|
| Time-multiplexed Structured Lighting using Gray Code |
| Direct Coding with a Colour Gradation Pattern |
| Direct Coding with a Colour Strip Pattern |

The interest here is to see whether a simpler single light projection coupled with a traditional stereo algorithm is competitive with a more complex coding scheme such as using a Gray Code constructed from multiple projections. A more detailed description of the presented Structured Lighting techniques can be found in [2].

**Photometric Stereo** An *Albedo Independent Approach* [6] with three light sources is used in this experiment. The focus in this paper is on comparing the gradient field integration component of Photometric Stereo. The algorithms were chosen to present both local and global algorithms. Global algorithms incorporate an optimisation process over the entire field and produce smoother results. The presented gradient field integration techniques are described in more detail in [15].

**Table 3. Photometric Techniques to Test**

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

## 4. Lab and Experiment Setup

A pair of *Canon EOS 10D* cameras with a measured focal length of 52 mm were used for high resolution image acquisition (an effective 6.3 megapixel resolution). The same cameras are used for both Binocular Stereo and Structured Lighting. The classic standard stereo geometry is used for the cameras. The baseline separation between the two cameras is 175 mm. The test subject is placed approximately 1200 mm horizontally away from the cameras. This system geometry was empirically chosen after extensive experimentation. The facial region was cut from the images to produce an $800 \times 700$ pixel region for comparison.

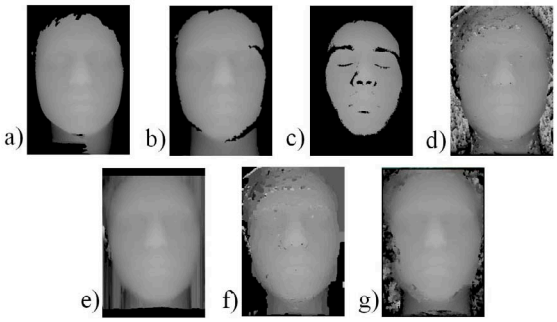


**Figure 1. Reconstruction examples: a) Ground Truth, b) Gray Code, c) FCV, d) SAD, e) SDPS, f) BVZ, g) CM.**

A cube shaped calibration object with 63 markings was used to calibrate the cameras.

An *Acer LCD Projector (model PL111)* was used to project structured light into the scene. The device is capable of projecting at a resolution of $800 \times 600$ pixels and has a focal length of $21.5 \leftrightarrow 28$ mm.

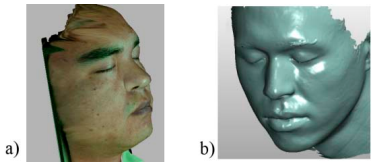


**Figure 2. a) Textured Gray Code and b) Ground Truth visualisations.**

The Photometric Stereo system geometry with three light sources is utilised [6] using $150W$ light bulbs and a *JVC KY-F55B* camera controlled automatically by a switching device. A calibration sphere is used to analytically determine the directions to the lights from an arbitrary scene origin.

A *Solutionix Rexcan 400* 3-D scanner was used to simultaneously obtain the ground truth data for each test subject (an example is shown in Fig. 2).

## 5. Experimental Results

Experiments were conducted on a Pentium 4 3.4 GHz machine with 2 Gigabytes of RAM. Comparisons are made between the resultant face reconstructions and a ground truth of the test subject. A set of 15 subjects were used for comparative analysis (this will be increased in the near future). Data alignment is conducted using a semi-automatic process involving 3-D object rigid transformations. The reconstruction accuracy is evaluated with the percentage of pixels with absolute depth errors less than two disparity units ($P_{<2}$), the mean ($e_{\mathrm{mn}}$) absolute pixel depth error, the standard deviation ($\sigma_e$) of errors, and the mean cosine error (MCE). Central differencing is used to estimate surface normals, and the MCE measures the quality of reconstruction of surface normals:

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

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 "$\bullet$" denotes the dot product of two vectors. The MCE indicates 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.

All presented algorithms were run with their given default parameter settings. However, it is notable that some tuning of parameters may give better accuracy for each algorithm.

**Table 4. Average accuracy on the database and running time for all tested techniques.**

| Method | $P_{<2}$, % | max | $e_{\mathrm{mn}}$ | $\sigma_e$ | MCE | Time, *sec* |
|---|---|---|---|---|---|---|
| **Gray Code** | 96.6 | 7.8 | 0.6 | 0.6 | 0.014 | 4.0 |
| **SDPS** | 88.5 | 13.0 | 1.0 | 0.9 | 0.092 | 6.0 |
| *SDPS + Gradation* | 89.8 | 12.7 | 1.0 | 1.0 | 0.111 | . |
| *SDPS + Strip* | 93.4 | 9.4 | 0.8 | 0.7 | 0.085 | . |
| **DPM** | 79.1 | 19.4 | 1.4 | 1.6 | 0.235 | 6.0 |
| *DPM + Gradation* | 84.3 | 13.3 | 1.2 | 1.2 | 0.246 | . |
| *DPM + Strip* | 92.4 | 12.6 | 0.8 | 0.8 | 0.138 | . |
| **BVZ** | 77.2 | 42.0 | 1.8 | 3.4 | 0.120 | 3517.0 |
| *BVZ + Gradation* | 82.9 | 30.5 | 1.3 | 1.5 | 0.092 | . |
| *BVZ + Strip* | 91.5 | 39.6 | 0.9 | 1.6 | 0.090 | . |
| **SAD** | 79.7 | 42.3 | 1.8 | 3.4 | 0.167 | 1.7 |
| *SAD + Gradation* | 84.9 | 32.0 | 1.2 | 1.7 | 0.158 | . |
| *SAD + Strip* | 93.1 | 35.1 | 0.8 | 1.3 | 0.089 | . |
| **BP** | 73.1 | 27.2 | 2.1 | 3.0 | 0.182 | 180.0 |
| *BP + Gradation* | 76.6 | 20.8 | 1.8 | 2.3 | 0.161 | . |
| *BP + Strip* | 88.5 | 17.6 | 1.0 | 1.2 | 0.162 | . |
| **CM** | 88.1 | 19.5 | 1.0 | 1.1 | 0.091 | 30.0 |
| *CM + Gradation* | 88.5 | 21.7 | 1.2 | 1.4 | 0.126 | . |
| *CM + Strip* | 92.1 | 20.7 | 0.9 | 1.1 | 0.098 | . |
| **PSM FCV** | 69.1 | 13.5 | 1.7 | 1.7 | 0.086 | 4.0 |
| **PSM Four-path** | 53.5 | 12.9 | 2.4 | 2.0 | 0.045 | 37.0 |
| **PSM Shapelet** | 71.2 | 11.6 | 1.7 | 1.7 | 0.036 | 153.0 |

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

## 6. Conclusion and Future Work

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).

The performance of a pure Gray Code approach is clearly ahead of other techniques. This is quantitively shown in its attainment of 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 [2] where PSM techniques usually fail.

Our comparative results differ from the universally accepted in Computer Vision ranking of stereo algorithms in [13] and the Middlebury Stereo Vision web page (www.middlebury.edu/stereo/).

It was found that global algorithms based on more complex optimisation techniques such as belief propagation (BP) and the graph minimum cut (BVZ) did not perform as well as expected for the case of human faces and relatively large disparity ranges. Here the accuracy of Dynamic Programming based algorithms is similar or even better than for the much more computationally complex Graph Cut algorithm.

Nonetheless, all tested algorithms show reconstruction errors that are not of a standard for direct usage in presenting 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 textuered did not show 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 textureless regions of the face which is why the projection of a colour strip pattern saw a marked improvement in reconstruction result.

Photometric stereo, although active in nature, is unable to recover true depth measurements due to the required gradient field integration step. None of the tested algorithms show performance comparable to the best offerings found in the other two techniques.

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.

Further investigation into the localisation of errors over the face will be conducted. The inclusion of more algorithms is being undertaken, especially for Binocular Stereo. Along with this, the creation of a larger database of test subjects is an ongoing and ever increasing project.

## References


[1] M. Chan, P. Delmas, G. Gimelfarb, 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, Sept. 2005*, LNCS 3691, pp. 740–747, 2005.

[2] P. Delmas, D. An, A. Woodward 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, 2005.

[3] R. Enciso, J. Li, D. Fidaleo, T.-Y. Kim, J.-Y. Noh, , and U. Neumann. Synthesis of 3d faces. In *Proc. Int. Workshop Digital and Computational Video*, Tampa, Florida, U.S.A, Dec. 1999.

[4] A. M. Haider and T. Kaneko. Automatic Reconstruction of 3D Human Face from CT and Color Photographs. *IEICE Trans. Information and Systems*, volume 12, pages 1287–1293, Sept. 1999.

[5] S. B. Kang and M. Jones. Appearance-Based Structure from Motion Using Linear Classes of 3-D Models. *Int. J. Computer Vision*, vol. 49, pp. 5–22, 2002.

[6] R. Klette and K. Schluns. *Computer Vision - Three-dimensional Data from Images*. Springer, 1998.

[7] 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.

[8] 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. Conf. Advanced Concepts for Intelligent Vision Systems (ACIVS'04), Brussels, Belgium, Sept. 2004*.

[9] W. S. Lee and N. Magnenat-Thalmann. Fast head modeling for animation. *Image and Vision Computing*, vol. 18(4), pp. 355–364, 2000.

[10] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proc. ACM SIGGRAPH'95 Conf.*, pp.55–62, 1995.

[11] Z. Liu, Z. Zhang, D. Adler, M. F. Cohen, E. Hanson, and Y. Shan. Robust and rapid generation of animated faces from video images: A model-based modeling approach. *Int. J. Computer Vision*, vol. 58(2), pp. 93–119, 2004.

[12] F. Pighin, R. Szeliski, and D. Salesin. Modeling and animating realistic faces from images. *Int. J. Computer Vision*, June 2002.

[13] 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.

[14] A. Woodward and P. Delmas. Towards a low cost realistic human face modelling and animation framework. In *Proc. Int. Conf. Image and Vision Computing New Zealand (IVCNZ'04), Akaroa, Christchurch, New Zealand, Nov. 2004*, pp. 11–16.

[15] A. 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 Conf., Dunedin, New Zealand, 2005*.

[16] Y. Zhang, T. Sim, and C. L. Tan. Rapid modeling of 3D faces for animation using an efficient adaptation algorithm. In *Proc. 2nd International Conference Computer Graphics and Interactive Techniques in Australasia and South East Asia, Singapore, June 2004*, pp. 173 – 181.

[17] P.F. Felzenszwalb, and D.P. Huttenlocher. Efficient Belief Propagation for Early Vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, June 2004*, IEEE CS Press: Los Alamitos, pp. 261–268, 2004.