3D Face Reconstruction from 2D Images
A Survey

W.N. Widanagamaachchi
University of Colombo School of Computing
35, Reid Avenue, Colombo 7, Sri Lanka.
wathsy31@gmail.com

A.T. Dharmaratne
University of Colombo School of Computing
35, Reid Avenue, Colombo 7, Sri Lanka.
atd@ucsc.cmb.ac.lk

Abstract

This paper surveys the topic of 3D face reconstruction using 2D images from a computer science perspective. Various approaches have been proposed as solutions for this problem but most have their limitations and drawbacks. Shape from shading, Shape from silhouettes, Shape from motion and Analysis by synthesis using morphable models are currently regarded as the main methods of attaining the facial information for reconstruction of its 3D counterpart. Though this topic has gained a lot of importance and popularity, a fully accurate facial reconstruction mechanism has not yet being identified due to the complexity and ambiguity involved. This paper discusses about the general approaches of 3D face reconstruction and their drawbacks. It concludes with an analysis of several implementations and some speculations about the future of 3D face reconstruction.

1 Introduction

The humans can perceive the 3D (3 Dimensional) shape of a 2D (2 Dimensional) image by just looking at it, even if the object in the image is completely new to the eye. The human brain plays a vital role in obtaining this 3D world through 2D images. After noticing the 2D image, the human eye signals the brain about the object through a nerve signal. After processing the nerve signal the brain creates the 3D shape of the 2D object. Appearance of the object, familiarity with shapes of similar 3D objects and other similar factors assist in creating the aforementioned 3D shape [10, 2]. Though this is an unconscious act for the humans, when tried to simulate with computers, efficient and effective ways have to be explored for identifying object features to assist the reconstruction of the 3D face. Thus it makes the area of 3D shape reconstruction from 2D images a complex and a problematic one.

The topic, 3D face reconstruction from 2D images has been derived and studied separately from the more general area of 3D shape reconstruction due to its depth and the complexity.

Techniques for attaining facial information for 3D reconstruction are broadly categorized into three, namely, pure image-based techniques, hybrid image-based techniques and 3D scanning techniques. The pure image-based techniques perform the reconstruction using only 2D images without estimating the real 3D structure. In hybrid image-based techniques both approximations and the data gained from images are used in the reconstruction process. The 3D scanning techniques have the capability to capture the complete 3D structure since scanned images provide both geometry and texture information of the face.

Human face is difficult to model even using normal 3D modeling software; hence the task of reconstructing them according to features gained from 2D images and making them realistic and accurate is, without doubt, even more intricate. The individual shape and variations in the human face, varying reflectance properties of the skin and actual depth estimation of face components add up to that intricacy [7]. Consequently, this topic has become one of the fundamental problems in computer vision at present [10].

The need for 3D face reconstruction has grown in applications like virtual reality simulations, plastic surgery simulations, face recognition, face morphing, 3D games, human computer interaction and animations [7, 6]. Though extensive research has being carried out, a fully accurate facial reconstruction mechanism has not yet being proposed [8].

The work in the early mechanisms of facial reconstruction focused only on producing realistic faces, however today, accurate reconstructions for facial plastic surgery simulations and fast and simpler reconstructions for 3D games have also become a necessity.

The rest of the report is organized as follows. Section 2 provides a detailed description about the general approaches for 3D face reconstruction from 2D images. In spite of the varied differentiations of implementation, there are some

978-0-7695-3456-5/08 $25.00 © 2008 IEEE
DOI 10.1109/DICTA.2008.83

preliminary steps which should be included in such a reconstruction process. Section 3 is devoted to describing those steps. The limitations and complications faced in a 3D face reconstruction are summarized in section 4. Though there are numerous reconstruction techniques, section 5 focuses only on a chosen few to highlight their divergent approaches. Finally, section 6 concludes with some speculations about the future of 3D face reconstruction from 2D images.

2 General Approaches

There are many approaches for reconstructing 3D faces but the choice of approach may vary according to the application for which the reconstruction is used. The general approaches are shape from shading, shape from silhouettes and shape from motion and analysis by synthesis using morphable models [1].

The most successful approach up-to-date is **analysis by synthesis** in which the parameters of the 3D statistical model are adjusted to increase the accuracy between the reconstructed face and the 2D face image. The errors in this approach are caused by 3D-2D alignment, shape differences, illumination differences and the quality of the dense correspondence among the 3D surfaces [1].

Despite the advances in depth estimation, **shape from shading** remains important because it overlooks most of the shortcomings of depth estimation. Algorithms for recovering shape from shading are generally considered to yield very good results in global minimization while local approaches are more erroneous but faster [7]. The Tsai-Shah algorithm which is used by Fanany et al. [7] is an example of the local approach.

A silhouette is an outline, shape or shadow of an object. Silhouettes provide accurate and robust data for reconstructions since they depend only on shape and pose of the object and are illumination-independent. These silhouettes, if extracted from the input images provide accurate data for the reconstruction process. Both Samaras et al. [14] approach and Lee et al. [12] approach use silhouette images to recover shape.

Just as humans use prior knowledge on similar objects to perceive 3D images, in a computer implementation a **database** and/or a **generic 3D face model** can be used as prior knowledge [10]. Normally face images and/or depth maps and texture information can be stored in databases. The input image and these stored images are compared, and the corresponding images are exploited in determining the facial components of the input image. The depth maps of these corresponding images assist in estimating the depth of the face components.

However it is very unlikely for the input image to contain an image of a face which resides in the database. Even if the input image contains such a face it can have different lighting and viewing conditions. Therefore techniques to exploit the stored images to produce novel 3D faces and techniques to reconstruct a face in different lighting conditions should be thoroughly explored.

Human face has a basic structure with features such as nose, mouth and eyes, but within these features there are minor differences which make a person unique. Researchers have designated around 150 **feature points** (figure 1) that can be used to capture these minor differences [4, 13]. Approaches which use these feature points can perform automatic or user driven feature point extractions.

As a result of the recent research conducted by Microsoft, [11] software was produced which automatically locates 83 feature points of the face, but the input image has some limitations. The image should be a frontal face, having a neutral expression and should be in normal illumination.

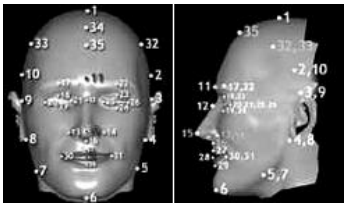

Figure 1. Feature Points. [13]

Since the details of the face are extracted from input images many considerations have to be made in deciding the number of the images required and the viewpoint of those images. Some argue that implementations based on multiple images are more liable to obtain accurate reconstructions since more data about the face can be grasped.

When comparing with other arbitrary viewpoint face images the frontal image captures all the face features. Due to this reason, most implementations based on single image require the image to be a frontal face with a neutral expression.

Birkbeck et al. [3] took an approach which rotates the person on a turntable to acquire a set of images from different viewpoints while Gong et al. [8] took images across views from minus 90 to plus 90 degrees at 10 degree increments. The camera is adjusted according to a magnetic sensor which is attached to the head.

In Birkbeck et al. [3] approach, all steps from image-capturing to 3D face reconstruction are performed through a GUI (Graphical User Interface) program. The shape is obtained by silhouettes and the texture is generated with the use of conformal mapping to reduce the distortion which occurs when 3D surfaces are flattened in to 2D space. At the time of rendering, the correct texture for each viewpoint is modulated from the textures.

Rasiwasia [13] took only two images into consideration - a frontal and a profile view. Since limitations in the input image's viewpoint cause inflexibility, researchers have currently focused on reconstructing faces from single 2D image where the image has no limitation in pose or expression and it can be taken in an arbitrary viewpoint. Guan's [9] approach provides useful groundwork in that region.

3 Steps in a regular 3D face reconstruction approach

After considering all these approaches, a set of general steps can be derived which will included in a regular 3D face reconstruction algorithm. The following is a list of the identified steps.

- Repairing the damaged areas (caused by noise, occlusion or shadows)

  The input image's condition might not always be satisfactory; they may be damaged or corrupted. Noise pixels of the image, if exist, might lead to inaccurate reconstructions. Shadows, poor lighting conditions and occlusions prevent accurate feature extraction of the face. Due to these reasons these damaged areas need to be eliminated prior to reconstruction.

- Face localization

  Few approaches like Rasiwasia's method [13] involve predefined restrictions in the input images. Although these restrictions introduce inflexibility, they reduce the complexity and preclude other face localization difficulties.

  Since input images in non-restricted approaches may contain other background elements apart from the human face, the face region should be identified and cropped. The distinctive color of the human skin can be used as a guide in identifying the face region. This process is labeled as face localization.

  In approaches where multiple images are being taken as input, each input image has to be cut and resized to obtain face regions. In addition, all these obtained image parts should be precisely aligned with each other.

- Facial component detection

  After the face region is isolated, the components of the face can be easily identified. Image-based techniques, silhouettes and feature points can be used to detect these facial components. In identifying these facial components, recognizing the two corners of the eyes, tip of the nose and the center and end points of the mouth would prove enough.

- Depth estimation

  For an accurate and realistic reconstruction, both location and depth of the facial features of the reconstructed face should be equivalent to the real face. Constructing the depth map of the input image will assist in depth estimation.

- 3D face reconstruction

  After face components' locations and depth are identified the 3D face can be reconstructed. A default 3D model can be deformed according to the real features to obtain the final 3D face. The texture should be mapped onto the 3D face. This is an intricate process since the texture information gained from 2D space has to be mapped onto a 3D space. Some approaches project the frontal image directly onto the 3D face but if the approach takes multiple input images these images can be warped into the texture space to generate a more realistic effect. The above mentioned Microsoft's approach [11] projects the frontal image directly onto the 3D face while Birkbeck et al. [3] warps the input images to the texture space.

4 Difficulties in 3D face reconstruction from 2D images

The uncertainty which lies in facial component detection can be eliminated by using multiple images but it might not always be possible to attain that many images. Even if multiple images are available, factors like noise, occlusion and shadows and/or lack of features in images might prevent the system from using them. To make the matters worse, multiple images might make the problem of time and effort even more obvious. The time issue is mainly caused by the preprocessing phase required.

As a result most researchers' attention has narrowed down to single image based 3D face reconstructions. One image of a face does not provide sufficient information for a 3D reconstruction, even if it's a frontal image. If the implementation has limitations in viewpoint, the input image may not even contain all the facial components.

Human face belongs to a particular class of similar objects. This class can be used in making inferences about the human face to assist in generating other views of the face in the aforementioned circumstance. A database which is maintained within the implementation can facilitate in making these inferences.

In maintaining a database the main dilemma lies in deciding the size of it. Unless the input 2D image's viewing

conditions are known in advance, images of each face taken under different lighting and viewing conditions have to be stored but large storage requirements, increased probability of false matching and slower reconstructions makes this option rather impractical. Basri and Hassner [2] presented a novel solution which answers this problem.

'Feature points' is a well-liked method for facial component detection but using countless feature points in the application can lead to inefficiencies in the computational time taken. Therefore approaches that involve a smaller number of feature points have gained recognition. Blanz et al. [4] approach is an example for such an approach. In recovering 3D facial information from multiple images the relationship between feature points in different viewpoints should be maintained.

5 Recent work

Blanz et al. [4] put forth a reconstruction approach based on a small set of feature points, a reference face and a database. The locations of the feature points are set in the reference face so that it can be used to automatically extract feature points from the input image. Additional feature points are used for texture reconstruction. The reconstruction is carried out by merging the stored shapes and textures in the database to correspond to the positions and gray values of the actual feature points.

Since 2D shape information and texture information are considered individually, the reconstructing process has two alternatives for the texture of the 3D face - the standard texture of the reference face or the true texture. A *'x by x'* mask is applied on each point and the mid value is obtained as texture information in the hope of reducing errors caused by noise. In experiments 22 shape reconstruction feature points, 3 texture reconstruction feature points and a '3 by 3' mask have being used.

The limitation that face should not have glasses, earrings or beard is a setback in this approach. The resolution of the images is limited to 256 x 256 pixels and colored images are converted to 8-bit gray level images.

Basri and Hassner [2] present a MATLAB-based solution with an underlying database which has an update mechanism. The images are partitioned into classes assuming that similar looking objects have similar shapes (e.g. fish, face) and the database was created by storing these images in the same class along with their depth maps. Since the input image's viewing conditions are not known in advance they have stored images of the same object with different viewing conditions in the database. Though this also results in an infinite example database the problems which arise with it are eliminated by the use of an update scheme. Starting with an initial seed of the database, it updates *on-the-fly* during processing in a way such that least used examples are replaced with more suitable 3D objects with better viewing conditions. As a result only a small relevant subset of the database is accessible to a user at any given time.

In performing the depth estimation of the face, parts of the image are compared with the image parts in the database to match the intensity patterns (figure 2). The found intensity patterns are taken as the initial guess for the face's depth and later a global optimization scheme is applied for depth refinement. When using a Pentium 4, 2.8GHz computer with 2GB RAM for a 200 x 150 pixel image via 12 example images at a given time, the running time of this application is around 40 long minutes.

The ability to handle a large database and being applicable to a variety of objects irrelevant of their viewing and lighting conditions makes this a successful approach.

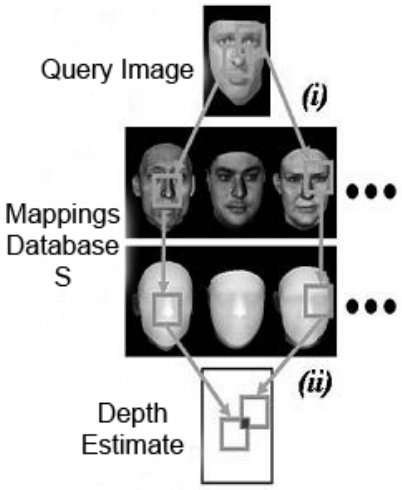

Figure 2. Visualization of the Process. [2]

Fanany et al. [7] present a neural-network learning scheme for 3D face reconstruction. This system can process the polygon's vertices parameter of the initial 3D shape based on depth maps of several images taken from multiple views. These depth maps are obtained by Tsai-Shah shape-from-shading (SFS) algorithm. An appropriate initial 3D shape should be selected in order to improve model resolution and learning stability. The texturing is performed by mapping the texture of face images onto the initial 3D shape.

The NN (Neural Network) scheme can store vertices of a 3D polygonal object shape. These vertices of the object in 3D space could be updated by the use of error back propagation after comparing the projected images with the real images. Since the NN could generate only flat projected polygonal models as its output, they have added a Gouraud smooth shading module to post-process the output of the NN. Hence the whole scheme is named Smooth Projected Polygon Representation Neural Network (SPPRNN). Ver-

tex, color and camera are the three parameters of the projected polygon representation NN.

The Tsai-Shah SFS algorithm processes both input images and NN output images in order to reconstruct the 3D face based on the depth maps. These depth maps are considered as partial 3D shapes rather than images.

In Samaras et al. [14] approach, 3D shape is extracted from Multi posed face images taken under arbitrary lighting and the reconstruction process uses silhouette images. The accuracy of this reconstruction process lies on the number and location of cameras used to capture the input images. A 3D face model is used as prior knowledge to assist in the reconstruction process.

The 3D face model is constructed from a set of 3D faces attained from 3D scanning technologies. The shape and pose parameters are estimated by minimizing the difference between the face model and input images. Later the illumination and spherical harmonic basis parameters are extracted from the recovered 3D shape.

Figure 3. Silhouette Extraction [14]

Rasiwasia [13] presents a simple and easily understood approach based on two orthogonal pictures - frontal view and profile view. The input images can be obtained by a stereo camera or a hand held camera but with the constraint of being in normal white light with a background which is free from any skin colored objects. 35 feature points and a generic model are used in this reconstruction process. The complete system is implemented using MATLAB.

The user is asked to indicate four specific points in each image - Eye, Nose, Mouth and Ear. The transformations for aligning the two images are calculated based on those points. When aligning, the images are scaled, rotated and translated till the frontal and profile images are in a horizontal line.

$$\theta = \sin^{-1}(A/(B^2 + C^2) - \tan^{-1}(C/B)) \qquad (1)$$

Theta in (1) is the angle by which the profile image needs to be rotated.

A = desired Y difference calculated from the ear and nose point in frontal image
B = actual X difference between the ear and nose in the profile image
C = actual Y difference between the ear and nose in the profile image

The distinctive color of the human skin is used in identifying the face region within the image. The (R, G, B) in the images is classified as skin if it satisfies the following conditions.

R $>$ 95 and G $>$ 40 and B $>$ 20 and
max$\{R, G, B\}$ - min$\{R, G, B\} > 15$ and
|R - G| > 15 and
R - G > 20 and R - B > 20

Figure 4. Skin Detection [13]

In extracting feature points, pure image based techniques are used. X and Y coordinates (X_f, Y_f) of a feature point can be obtained from the frontal image while the Z coordinate along with the Y coordinate (Y_p, Z_p) can be attained from the profile image. Since images are aligned, both these Y coordinates are approximately the same.

So the final feature point coordinates can be achieved for all the 35 feature points by using (2).

(X_f[i], (Y_f[i]+Y_p[i])/2, Z_p[i]) where i=1, 2, ..., 35 (2)

Figure 5. Generic Eye Template [13]

A template matching algorithm (figure 5) and prewitt operator is used in extracting the feature points of the eye from the frontal image while horizontal and vertical histograms are used to detect the location of the mouth. After the feature points of the eyes have being extracted a rectangular region (figure 6) is cropped out from the frontal face. This rectangular region's left and right boundaries are the farthest point of the eyes and the upper boundary is the lower part of the eyes. The horizontal histogram is drawn on this cropped region and the first peak from the top after a certain threshold is used to identify the location of the mouth.

The center of the mouth is identified by drawing a vertical histogram in this localized mouth region.

Figure 6. Rectangular Region and the Horizontal Histogram for Mouth [13]

Though all the 35 features can be automatically identified, at the end of the extraction process, this method offers the capability for any user modifications if required. These feature points that are found are then used to deform the generic model. This deformation is done in two steps - Globally and Locally. Finally the texturing of the face is performed using the frontal image in a manner that actual features in the reconstructed face overlap with the features in the frontal image.

The following image (figure 7) presents some reconstructed faces of this approach.

Figure 7. Example Reconstructed Faces [13]

Recently an automatic reconstruction based on a 3D generic face and a single image (irrelevant of pose and expression) was presented by Guan [9]. The only condition required in the image of the face was for the head rotation to be in the interval +30 degrees to -30 degrees. This method is said to reconstruct 3D faces with standard and low cost equipments. The features extracted from the images serve as geometric information which helps in deforming the 3D generic face. The feature points are detected by using Euclidean angles. It is assumed that the head is not rotated with respect to the X axis.

The texturing of the face (figure 8) is performed by orthogonally projecting the 2D images onto the 3D face. When the 2D image is orthogonally projected to form the texture, some vertices contain no corresponding color since they are occluded. Those vertices generate blank areas in the texture. As a result a thin-plate relaxation method is used in interpolating those blank areas with known colors.

Gong et al. [8] put forth a multi-view nonlinear shape

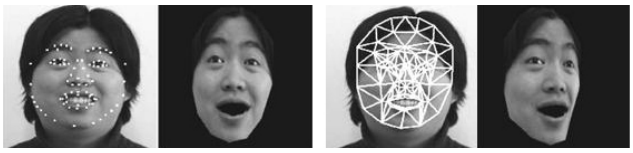

Figure 8. 3D face reconstruction with an open mouth expression [9]

model which is 2D view-dependent but has no reference to 3D structures. They have used a Kernel PCA (Principal Components Analysis) based on Support Vector Machines for nonlinear shape model transformation.

This method has found remedies for two main drawbacks which occurred because of the large pose variations of human face. Nonlinear shape transformations across views using Kernel PCA based on support vector machines is used to address the first problem which is highly nonlinear shape variations across views. The second drawback of unreliable relationships among feature points across views (based solely on local gray-levels) was addressed by improving a nonlinear 2D active shape model with pose constraint.

Figure 9. Shapes fitted to Images of an unknown face across Views using the view-context based nonlinear ASM (Active Shape Models) [8]

Darrell et al. [5] present a method based on cubical ray projection. This algorithm uses a novel data structure named 'linked voxel space'. A voxel space is used to maintain an intermediate representation of the final 3D model. Since connectivity of the meshes cannot be represented and converting a volumetric model to a mesh is difficult, a linked voxel space is used instead of a voxel space.

First the 3D views obtained from stereo cameras are registered based on a gradient-based registration algorithm. The result of this registration is a 3D mesh where each vertex corresponds to a valid image pixel. The location of each vertex in the mesh is calculated and mapped in to a voxel. This voxel space is reduced using a cubic ray projection merging algorithm. This reduction is done by merging the voxels which fall on the same projection ray.

Since this method uses stereo cameras to get the synchronized range and intensity 3D views texture alignment might not be a necessity.

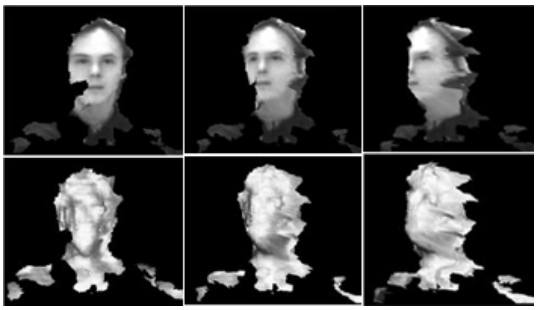

Figure 10. Final 3D mesh viewed from different directions [5]

6 Conclusion

The 2D image of a face is very sensitive to changes in head pose and expressions so a successful reconstruction approach should be able to extract these face details in spite of these changes. Approaches based on silhouettes and prior knowledge can be advantageous in addressing this problem. When reconstructing 3D faces from 2D images the key source of information is the intensity based features and landmarks of the image. But intensity alone is not enough in case of low intensity, noise, occlusion, illumination variations and/or shadows being present in the input images. The anatomical landmarks are argued to be a more accurate source of information, but they are rather thin and difficult to locate.

Most traditional face reconstructions require a special setup, expensive hardware, predefined conditions and/or manual labor which make them impractical for use in general applications. Though recent approaches have triumphed over some of these setbacks, quality and speed are not still up to the expected levels. More realistic 3D character modeling software could be used in reconstructing the final 3D face or the default 3D model can be created from such software.

Strategies like supervised learning and unsupervised learning in neural networks can be applied in facial component identification. Fuzzy systems can be used in feature extraction processes for a more fruitful result.

Prior knowledge of a face in different viewing and lighting conditions can be stored in the database with efficient update schemes which would eliminate the uncertainty involved in reconstruction from a single arbitrary image. The recent successful approaches should be continued and refined to adhere to the changing requirements of the modern society. Limitations like not having a beard, not wearing earrings and glasses should also be eliminated.

Most present reconstructions are limited to reconstruct a face with just the front area. These reconstructions should be extended to reconstruct a face with realistic hair and ears. When an arbitrary image is given the system should be able to draw out necessary inferences to obtain other views of the face.

The topic 3D face reconstruction from 2D images has retained its significance in the computer world and with the recent development; applications like human expression analysis and video conferencing have been added to the long list of its applications. Virtual hair and beauty salons is one future application where 3D reconstructed faces will prove to be valuable. Having the opportunity of viewing the aftermath of a haircut or a facial before even getting it and sometimes even viewing the face of a long-gone person is without doubt a priceless reward. The 3D face reconstruction can be extended to produce aging software which have the capability to produce younger or the older face of the input image.

References

[1] S. Amin and D. Gillies. Analysis of 3d face reconstruction. In *Proceedings of the 14th IEEE International Conference on Image Analysis and Processing*, 2007.

[2] R. Basri and T. Hassner. Example based 3d reconstruction from single 2d images.

[3] N. Birkbeck, D. Cobzas, M. Jagersand, A. Rachmielowski, and K. Yerex. Quick and easy capture of 3d object models from 2d images.

[4] V. Blanz, B. Hwang, S. Lee, and T. Vetter. Face reconstruction from a small number of feature points.

[5] T. Darrell, L. Morency, and A. Rahimi. Fast 3d model acquisition from stereo images.

[6] E. Elyan and H. Ugail. Reconstruction of 3d human facial images using partial differential equations. *Journal of Computers*, 2(8), 2007.

[7] M. Fanany, I. Kumazawa, and M. Ohno. Face reconstruction from shading using smooth projected polygon representation nn.

[8] S. Gong, A. Psarrou, and S. Romdhani. A multi-view nonlinear active shape model using kernel pca. *BMVC99*, pages 483–492.

[9] Y. Guan. Automatic 3d face reconstruction based on single 2d image. In *Proceedings of the IEEE International Conference on Multimedia and Ubiquitous Engineering*, 2007.

[10] F. Han and S. Zhu. Bayesian reconstruction of 3d shapes and scenes from a single image.

[11] Y. Hu, D. Jiang, S. Yan, H. Zhang, and L. Zhang. Automatic 3d reconstruction for face recognition. *Journal of Pattern Recognition*.

[12] J. Lee, R. Machiraju, B. Moghaddam, and H. Pfister. Silhouette-based 3d face shape recovery. *Graphics Interface*, 2003.

[13] N. Rasiwasia. The avatar: 3-d face reconstruction from two orthogonal pictures with application to facial makeover.

[14] D. Samaras, S. Wang, and L. Zhang. Face reconstruction across different poses and arbitrary illumination conditions. *AVBPA, LNCS*, pages 91–101, 2005.