Face Recognition and Computer Graphics for

Modelling Expressive Faces in 3D

by

Tufool Al-Nuaimi

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

May 26, 2006
[February 2007]

Author_____
Tufool Al-Nuaimi
Department of Electrical Engineering and Computer Science
May 26, 2006

Certified by_____
Judith Barry
Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Face Recognition and Computer Graphics for

Modeling Expressive Faces in 3D

by

Tufool Al-Nuaimi

Submitted to the Department of Electrical Engineering and Computer Science

May 26, 2006

In Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

This thesis addresses the problem of the lack of verisimilitude in animation. Since computer vision has been aimed at creating photo-realistic representations of environments and face recognition creates replicas of faces for recognition purposes, we research face recognition techniques to produce photo-realistic models of expressive faces that could be further developed and applied in animation. We use two methods that are commonly used in face recognition to gather information about the subject: 3D scanners and multiple 2D images. For the latter method, Maya is used for modeling.

Both methods produced accurate 3D models for a neutral face, but Maya allowed us to manually build 3D models and was therefore more successful in creating exaggerated facial expressions.

Thesis Supervisor: Judith Barry
Title: Research Fellow at The Center for Advanced Visual Studies

To mom and dad

Contents

List of Figures

Chapter 1

Introduction

In a world where we see and understand objects in three-dimensions, it is sometimes difficult to view details of an object in a two-dimensional representation such as a photograph or a drawing. The need for more detail from a model and a more realistic and natural representation is what has driven 3D modeling to progress rapidly and to become more available to consumers.

3D programs such as Maya, CATIA and 3D Studio Max, are applied in a wide area of disciplines; most notably in architecture, engineering and industrial design. These programs are tools for visualization as they allow users to model prototypes and analyze designs. Some of these programs have also been applied to art practices in the forms of animation and video games for example.

While the 3D programs mentioned above are used to create objects, computer vision can be thought of as the opposite: it is the study of how computers are able to perceive existing images using cameras and sensors then extract and process, from these images, specific data in order to portray them in 3D.

Although computer vision is applied to many fields, such as 3D laser guided surgery in medicine, face recognition, and 3D teleconferencing; it remains an area of

study that is still undergoing research. The limitations of computer vision arise from the lack of understanding of how the mind observes and registers objects, and due to errors in geometrical measurement of the object, especially when the object in question is small and with an unknown geometry, like hair.

It is not unusual to see that, in this computer age, modern art can be intertwined with, and can sometimes depend on, more technical disciplines. A common example is producing special effects for movies, which requires extensive computer programming in order to build a virtual world that can convey the verisimilitude of the 'real' world. Moreover, beyond computing, there are many other disciplines that effect art. An example is neuroscience where a realization of how the human brain registers objects can have positive impact when modeling "artistic" objects.

Acknowledging that technical disciplines influence the arts, this research explores the methods that are widely used in machine learning, and in particular, face recognition and applies them to computer graphics in order to investigate new methods for producing detailed animation.

The final goal is to deliver a product that can be further developed artistically. In particular, we want to model several photo-realistic faces in 3D with varying expressions that will be later fused together, using Maya, such that they portray some kind of visual effect, like frozen motion, for example.

Computer vision is a discipline where machines use various algorithms to interpret scenes and objects, and extract from them sufficient data about geometry and texture in order to reconstruct them in 3D. Face recognition, a branch of computer vision, is a narrower research area where the aim is for a computer to recognize a face.

Face recognition was developed to serve different disciplines like forensic science and security. Although older methods for security are still used, like finger printing and iris scans, recognizing a face--from a surveillance camera for example—offers a less intrusive security method. Since the techniques used for face recognition entail acquiring extensive detail from an object, and then reconstructing it such that it resembles the real face, they apply directly to the animation of a realistic face.

Chapter Two introduces face recognition by explaining how it has progressed throughout the years. More formally, the beginning of face recognition used different methodologies that reconstructed 2D images. Due to the weaknesses of 2D images, like their lack of sufficient geometric detail and intolerance to facial hair, and variations in facial orientation and makeup, they were replaced with 3D models which offer more flexibility and overcome the limitations of the prior method.

Chapter Two also discusses the evolution of animation from 2D to photo-realistic 3D, and the positive impact that computer vision can offer animation and graphics.

Chapter Three explores various aspects of face recognition that directly effect animation. The two main stages of face recognition are gathering geometric and texture data from an object and then reconstructing the object in 3D. We investigate data gathering methods like using still images, video and 3D scanners, and explore their advantages and weaknesses

Since the main goal is modeling a 3D face, we also explore (in chapter 3) the limitations of face recognition that would directly affect our results--like its lack of ability to recognize expressive faces, and the difficulty of modeling hair--in order to adjust these flaws using computer graphic techniques.

Chapter Four outlines the different experiments that we performed in order to model a face in 3D.We started by using multiflash strobe lights to investigate motion. We then used video and 3D scanners to gather necessary geometric data. The images that were gathered from video were imported into Maya to model a face in 3D, which was done in collaboration with Janet Tremblay from the Art Institute of Boston, under the guidance of Alex Hart, her Maya professor.

This chapter also includes experimental results, and a discussion of the advantages and disadvantages of the different methods that were used. Although the thesis of this project explores how computer vision has a positive impact on animation, a brief argument of how the opposite could also be true (using a method we employed when modeling the face) is also included in this chapter.

Chapter Two

Background

Being able to interpret details of scenes and objects, and reconstruct them in 3D has been a problem of long tradition in both computer vision and in animation. Although the applications differ in the two paradigms, the requirements are the same: producing photo-realistic images.

Since computers can gather intricate geometrical and textural data from objects, which cannot be inferred without such aid, computer vision has found its way into many disciplines including medicine, forensics, security and engineering.

Similarly, the goal of 3D representations in animation is creating fictional or non-fictional environments which simulate an experience that is analogous to a reality; and this is why animation can benefit from computer graphics.

2.1 A Brief History of Computer Vision and Face Recognition

Computer vision is a process through which a computer visualizes real world images and recognizes features and patterns in order to produce data or information about a scene. Computer vision is a relatively new science; the beginning stages took the form of recognizing patterns from 2D images. Although this area is still being investigated today,

most computer vision research has progressed to reconstruction of a 3D scene from either multiple 2D images or from 3D point clouds.

Face recognition is an application in computer vision that allows a computer to identify a person by dissecting the general facial geometry in order to extract features and recognize patterns. Although face recognition may find its way to many applications, its most common purposes are for surveillance, security and forensic research. Older identification methods that are still readily used today are fingerprinting, iris scans and utilizing identification documents such as driver's licenses and passports. Although these methods are effective, the need for a non-intrusive approach for identification has made face recognition a popular research area.

Many different algorithms and techniques exist today for the purpose of pattern recognition. Earlier techniques compared 2D images, but these methods are unreliable if a facial expression or head orientation do not match those of the images in the database, or in the presence of facial hair or makeup [1]. Modern methods use 3D models; many of which have been successful in identifying a scene, an object or a person accurately, but only under controlled conditions. Some of these conditions will be discussed in chapters 3 and 4 in relation to the face modeling experiment.

Another obstacle that mars the performance of recognition is the lack of thorough understanding about how the brain perceives different scenes. It is therefore difficult to implement algorithms that would allow machines to identify scenes in manner analogous to human vision [2].

2.2 Impact of Computer Vision on Graphics and Animation

2.2.1 Progression in animation: achieving virtual reality

Throughout the last few years, special effects in movies and video games have progressed dramatically; more specifically, earlier animations contained special effects that were recognizable but not necessarily convincing. The need for great degrees of verisimilitude fueled the development of 3D animation and enhanced the quality of special effects.

Although 2D animation has also evolved and is now computer generated, it has not achieved the realism that 3D animation offers as it cannot easily re-produce the illusion of space--it exists solely on the x and y axes; even if it is sometimes able, through the use of perspective within a two dimensional picture plane, to give the illusion of depth. In contrast, 3D animation produces depth as it adds the possibility of utilizing the z-axis, along side of the x and y axes.

Animation evolved from 2D to 3D; but the first '3-dimensional' movie was created using stereoscopic images which portray depth, hence giving the illusion of 3D. Invented in the early 1980's by Sir Charles Wheatstone, stereoscopy utilizes an approach that replicates animate vision; humans and animals are able to perceive depth due to the distance of approximately 65mm between the eyes. Similarly, stereoscopy uses a camera with two built-in lenses that are approximately 65mm apart to capture two images of the same scene, thus conveying more information than that from a single image and providing a sense of depth. For those reasons, stereoscopic images have been commonly used in face recognition.

Modeling software like Maya and 3ds Max have allowed animation to progress even further and produced *real* 3D models. A common example of this is in present video games, where a user is able to navigate the virtual environment such that the 3D models can be viewed at any angle.

2.2.2 Enhancing quality of computer graphics using computer vision techniques

3-Dimensional programs such as Maya, CATIA, Rhino, AutoCAD have been applied to a wide area of disciplines such as architecture, engineering and industrial design; and provide a tool for visualization as well as for producing prototypes, models, and analysis of designs. These programs are also applied within art-world practices in the forms of graphics, animation, video games and sculptures, and have been successful in achieving photo-realism.

In computer vision, extensive research has been aimed towards finding more effective solutions for computers to estimate and extract information about objects and scenes via images. These methods use algorithms that allow computers to understand human physiology, like gestures, expressions and movements from images, and then reconstruct them in 3D, thus achieving photo-realism. The same requirement--achieving realism—exists in computer graphics, and thus integrating the two disciplines can help animation achieve a greater degree of verisimilitude.

Since our research is geared towards modeling a realistic face in 3D, we incorporate face recognition techniques that would help in producing this. The experiment procedure and results are detailed in chapter four.

Chapter Three

Exploring Face Recognition

The first step in face recognition is gathering information about an object or a scene. Various gathering methods exist, like using 3D scanners or still photography. But due to the pros and cons of each, there is still a debate about which method is the best for this purpose.

The second step is creating a model that replicates a face using various algorithms, and then using the model to identify a person from a database of images. Many of these methods have been successful in identification with a low error-probability, but only under controlled external conditions. Examples of variables that affect face recognition are intensity of illumination and type of facial expression.

The rest of this chapter will discuss these two steps in more detail.

3.1 Data Gathering Methods in Face Recognition

The Face Recognition Grand Challenge (FRGC) experimented with six different pattern recognition techniques that are of common practice today [3], and formed conclusions about which methods produced better results. Of the six experiments, four that we investigate further in our experiment are:

1. High resolution, still frontal images with uniform illumination and background.
2. High resolution, still frontal images with varying illumination and background.
3. Multiple, still images of a face with varying orientations.
4. 3-dimensional scans of an object.

Based on their experimental results, Philips *et al* concluded that a 3D image does provide more information on the geometry of a face, but a single high resolution image and multiple high-resolution images performed better in these face recognition experiments. Moreover, Philips *et al* stress that their experimental results are not final or ultimate, and that results may vary depending on the dataset or experiment type [3].

3.1.1 Advantages and Disadvantges of 3D Scanners

3-Dimensional scanning contributes to providing virtual, photo-realistic reproductions of real objects; and is valuable in countless industries. Some applications of the 3D scanner are the production of archives to document antiques, historical artifacts and monuments; fabrication of prototypes and models in different industries such as medicine, manufacturing, engineering and architecture; animation and computer graphics; and, because of its ability to capture and break-down substantial detail of various objects, 3D scanners are very useful for reverse engineering.

Scanners can be separated into three major categories [4]: contact, passive and active. A contact scanner, like the name suggests, glides over the object to collect geometric information of x,y and z co-ordinates. Although it is effective in modeling objects, some of its disadvantages are that it may cause damage to fragile objects, that it is labor-

intensive since it requires a person to slide the scanner probe over the object, and that it produces less detail due to the size of the probe.

A passive scanner gathers point clouds[1] by detecting ambient light reflected off the object. Although this is a cheaper method of scanning when compared to an active scanner [4], it is less accurate, especially when the geometric shape of the object is more complex or when there is insufficient ambient illumination.

An active scanner has a built in light source--usually a laser beam--that is directed onto an object using a rotating, galvanometric mirror. This radiation is then detected using a sensor, and the co-ordinates are calculated using triangulation[2].

Although the three scanners are useful for different purposes; active scanners, like the Konika Minolta used in the FRGC experiments, have the advantage of being faster and more accurate than a contact scanner and would not require as much ambient light as the passive scanner. Thus, the active scanner is the ideal choice for modeling a face in 3D.

Because of their high accuracy in capturing geometrical and texture detail of scenes, 3D scanners are making their way to many different industries and could replace older methods that have been used. For example, instead of taking pictures of artifacts and models in museums for archiving purposes, a 3D scanner could be used, which would contain more informative detail about the object.

---

[1] Point clouds are sets containing three-dimensional points which represent the x, y and z coordinates of an object and contain all the geometric information about the object's surface.

[2] Triangulation is a method which uses trigonometry to identify an object's position. In 3D scanning, a light beam is directed onto an object, and then reflected from the object onto the receiver. The emitted and reflected beam form two sides of a triangle. The third side is the distance between the emitter and the receiver. Calculation of the object's coordinates can be done straightforwardly since the length of the triangle's third side is known, and the angles at which the beam is emitted and received are also known.

Although a 3D model does provide more data than a 2D image, there is still some debate on whether it is the best method for face recognition for various reasons:

1. The Konika Minolta website states that the laser registers coordinates of a particular point with an error of 0.1 mm. This means that fine detail such as human hair, of diameter in the order of micrometers, or fine lines and wrinkles cannot be modeled accurately. A reason could be a result of 3D triangulation. In his research, Chou states that errors always result from 3D triangulation because when a sensor detects the laser, the two beams do not intersect at the exact point on the object, but at a point slightly farther away [5]. Although this does decrease the accuracy of a model, a 3D scan of an object still contains more information than other imaging methods because of the availability of the z-axis .

2. Although an active scanner, unlike a passive scanner, does not depend on ambient light because it has a built-in laser light, it is still very sensitive to illumination. An active scanner is unable to scan transparent objects and performs less effectively when the object being scanned is darker because the detector cannot 'see' the laser beam. For similar reasons, if there are shadows on a person's face due lack of illumination, the end result—the 3D model—will contain holes, which are regions in the model that contain no information. Chapter four shows the results of face models from 3D scans.

3. Scanners are also very sensitive to motion; if a subject moves during the process, the point clouds which represent different features would be translated depending on the movement, which results to a facial deformation.

4. Following on from the third disadvantage, there is a time delay in acquisition of data between the shape channel and the texture channel of the scanner. Therefore, motion during scanning would result to a lower-resolution or blurry image [3].

3.1.2 Advantage and Disadvantages of Still 2D images

An argument that disputes the results of the FRGC experiment but, at the same time, supports their claim that the results are not decisive or final is an example that could occur in a real-life situation. In this example an image that is extracted from a surveillance camera shows a face that is tilted downwards—thus not a frontal, high-resolution image as in the FRGC experiments. Would a computer be able to identify the person from a series of frontal images? Or would a 3D model that could be rotated and viewed at different angles be a better fit for this scenario?

A single high-resolution 2D picture is useful for face recognition if the images in the database are of almost identical poses. Most commonly, these data base images are frontal—also called mug shots—and to be recognized accurately, the 2D picture

should be of matching orientation such that pattern recognition algorithms would have enough information for comparison.

Another method is to use a series of 2D pictures for comparison. This would provide more information than a single image and could therefore be more useful for face recognition. On the other hand, these pictures may still lack the information required for successful recognition. Other complications include makeup and facial hair, which make it difficult to infer the facial geometry and thus hinder the performance of recognition algorithms.

To overcome the problems associated with a 3D scanner, and the lack of information available from a single 2D image, Blanz *et al* suggest a method of face recognition from non-frontal images [4]. After capturing a series of non-frontal images of a person, Poser—a computer graphics software—is used to reconstruct the images into a 3D model. A 2D front view can then be extracted from the model for comparison with mug-shots in the data-base.

Multiple 2D images still lack a third dimension, and although they may provide, to some degree, a sense of depth, some inference must be made to model this object in 3D. This problem—depth of stereo from 2D images—is a long-standing problem in computer vision[3].

In our experiment we used a method that is somewhat similar to Blanz's, where we reconstruct a 3D face from multiple still images using Maya: a 3D modeling software. We found that multiple images provide sufficient information about the 3D

[3] Section 2.1.1 explains stereoscopy.

geometry of a face. And that, together with some understanding of facial physiology allowed us to produce an accurate 3D model.

3.2 The Difficulty of Recognizing Expressions

A popular research topic in face recognition is the ability to accurately identify an expressive face from a database of neutral faces. This topic is also a problem with long history in face recognition. The reason for this is similar to the ones associated with recognizing different orientations and poses, discussed in section 3.1.2.

The ideal choice for face recognition would be a neutral face, but in real-life situations, it is difficult to control that factor. In a study performed by Yacoob et al, where different facial expressions were tested for the purpose of face recognition, it was found that a natural smile is the choice that results to the least error [6]. This results from the proximity in facial appearance and muscle structure of a smiling face to a neutral face (when compared to other expressions).

A face goes through numerous facial expressions each day, changing from one moment to another. We are able to understand an emotion that is attached to a facial expression because it comes naturally. This is because of how an emotion is

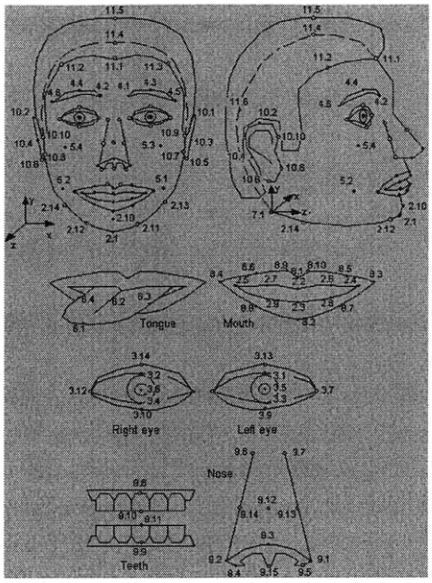


Figure 1: Neutral face containing 84 feature points as defined by Mpeg-4.

incorporated within changes in features, such as positioning of the eyes or shape of the mouth. Our mind is able to dissect an expression, understand it, and associate it with a person. But how does our mind perceive these expressions?

Cognitive neuroscientists do not have a complete understanding of how the human mind sees expressions, and it is therefore unfeasible to ascribe this method to machine learning. Until then, different methods are being applied to improve the performance of machines when identifying expressive faces.

Since the cause of the problem is that the face being identified (or 3D model) is expressive and the database images are neutral, the two obvious methods to solve this problem are to either change the database image such that it matches the expression or to change the 3D model such that it has a neutral expression[4] [7].

[4] These methods are not unique, as others are stated in literature.

In computer vision today, expressions are classified into six categories: happy, sad, surprise, disgust, anger and fear. Humans can categorize these expressions partly due to the geometry of the face and position of extremities; for example on a sad face, the corners of the eyes and lips are slightly tilted downwards. An *a priori* knowledge of the emotion associated with the face allows one to deform the face to resemble a neutral one, where a neutral face is pre-defined by the Mpeg-4 standard shown in Figure 1[8].

The method that Ramachandran *et al* employed to reduce error during detection was to convert an expressive face to a neutral one [9]. In their research, they inferred that only the positions and shapes of the jaws and lips change significantly while other features undergo minor variations during changes in expression. Therefore, using a piecewise method, the polygon meshes associated with the jaws and lips are normalized to resemble a neutral face.

In our experiment we used a similar method; building an accurate 3D model consisting of high-polygon meshes for the features. Furthermore, we introduced Bezier curves which replicate facial muscles such that the face can deform in a realistic manner; thus, giving us the ability to change an expressive face to a neutral one. The details of adding Bezier curves are in Chapter 4.

Chapter Four

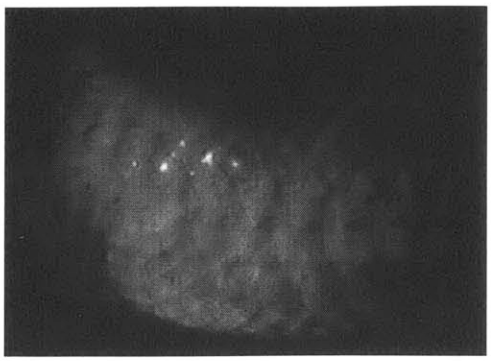
Experiment

4.1 Investigating motion using techniques from Muybridge and Edgerton.

The first phase of the experiment investigated various photography techniques aimed at capturing moving objects. Eadweard Muybridge and Harold Edgerton, separated by about one hundred years, were experts in capturing motion using photography. Their efforts in this field were aimed at freezing motion to study and portray detail that could not be seen with the naked eye alone.

A famous series of photographs by Muybridge was the running horse, captured by set of 12 cameras with high-speed shutters, aligned along a horizontal line. A circuit connected each camera to the horse-track, and a running horse triggered consecutive cameras by closing one circuit at a time. Since the pictures taken were proportional to the horse's speed, they portrayed realistic but extremely slow motion.

Methods that are analogous to those employed by Muybridge recently introduced a new class of special effects in films. The "Matrix effect" is an example of how Muybridge's techniques have progressed to create a 3D dramatic effect in movies. In this example, a camera array is placed along a circular track around a moving object, and consecutive cameras are triggered automatically with a short time delay between each pair such that they slow down or freeze a 360-degree action shot.



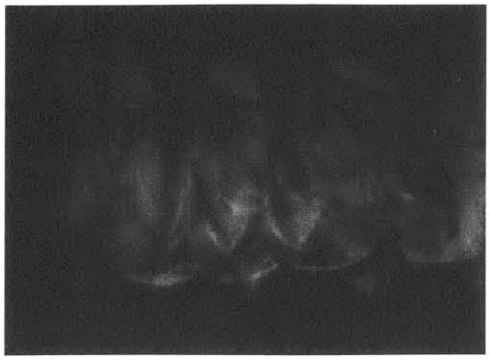


Figure 2a **Figure 2b**

Figure 2a: a multiflash strobe image of a rotating head through 180 degrees taken at 30 flashes per second with a shutter time of 1.5 seconds.
Figure 2b: a multiflash strobe image of a moving heat taken at 10 flashes per second with a shutter time of 1.5 seconds.

Decades after Muybridge's era, Edgerton used stroboscopy to freeze or slow down fast motion. A famous example is the frozen photograph of a bullet upon impact with an apple, which was taken using a microflash[5].

At the Edgerton lab, we used a multiflash strobe to capture a series of images that depict a 180-degree rotation of a moving head for the purpose of modeling it in 3D. The experiment was performed in a dark room, where the only source of light was the multiflash strobe. In order to provide the subject with sufficient illumination, the strobe was placed at a distance of 10 feet away from the subject, and a camera was placed 8 feet away from the subject. The experiment was repeated several times, using different flash frequencies and various types of motion. Figure 2 shows the results.

In Figure 2a, the flash rate was too high, causing superimposed images that minimize the definition of features and facial extremities that are vital for modeling in 3D. In addition, since the position of a rotating head does not vary significantly, superimposition

[5] A microflash unit has a short, intense flash that illuminates a moving object for period of time that is in the order of microseconds; and is therefore useful for very fast motion.

is hard to avoid. Figure 2b fixes the shortcomings of the previous method by decreasing the flash rate and increasing the range of motion. Regardless of the improvements achieved with the second method, we could not use it for 3D modeling for the following reasons:

1. Although the multiflash strobe we used had high-intensity illumination, the short bursts of light did not provide the illumination required to prevent the images from appearing "ghost-like". A wider aperture would allow more light to reach the film, which would have slightly improved the appearance of the images.

2. There is a time limit of up to 1.5 seconds for each sequence of flashes on the strobe used for this experiment, which makes it difficult for a subject to move through a 180 degrees but increasing flash-time would not improve results as it would cause stationary objects (like the background) to be overexposed.

Although the mulitflash photographs are not suitable for modeling a 3D face for the reasons mentioned above, they do portray the general appearance of a rotating head. In order to capture clearer and more detailed images, video was used.

4.2 Using video to capture motion

When reconstructing a model from a real object, the first step is gathering necessary information. In computer vision, there has always been a debate between the best approach for collecting this information. Although using video to capture motion has its disadvantages, like its low resolution when compared to high-resolution digital cameras, this method was chosen as the preliminary experiment because it portrays motion and changes over time, which is of major importance for capturing detail. More specifically, video will give us an insight about how facial expressions and features change from one frame to another and will allow us to examine the kinematics of a moving head, which is a feature that is not available when using a 3D scanner. In addition, multiple images are useful in gathering sufficient information about the features of an object [10].

The experiment was performed by setting a video camera on a tripod and adjusting it such that the moving face is always within the frame. The person then rotates through 360 degrees using a rotating chair that moves smoothly at a speed that is almost steady. Because the chair is spun manually, the exact rate of rotation is unavailable *a priori*, but can be estimated using Equation 1.

$$DPS = \frac{frame\_rate}{frames / rotation} \qquad \textbf{Equation 1}$$

Where DPS is degrees (of rotation) per second, frame rate is that of the video camera and frames/rotation can be found by counting: after the video is de-compiled into frames, the number of frames that make up a complete, 360° rotation is the frames/rotation.

The angle between frames is also unknown *a priori* but can be calculated using Equation 2.

$$\phi = \frac{1}{frame\_rate} \times DPS \qquad \textbf{Equation 2}$$

The experiment was repeated several times, each time with a different facial expression and head orientation. Although some of the expressions were amongst the common six used in computer vision: happy, sad, anger, disgust, surprise and fear, others were exaggerated. A selection of these image sequences is shown in Figures 3, 4 and 5; where Figure 3 shows a sequence of seven images of a neutral face with approximately 25 degrees of separation between consecutive images, Figure 4 shows a sequence with an exaggerated facial expression and Figure 5 shows a tilted face, also with an exaggerated facial expression.

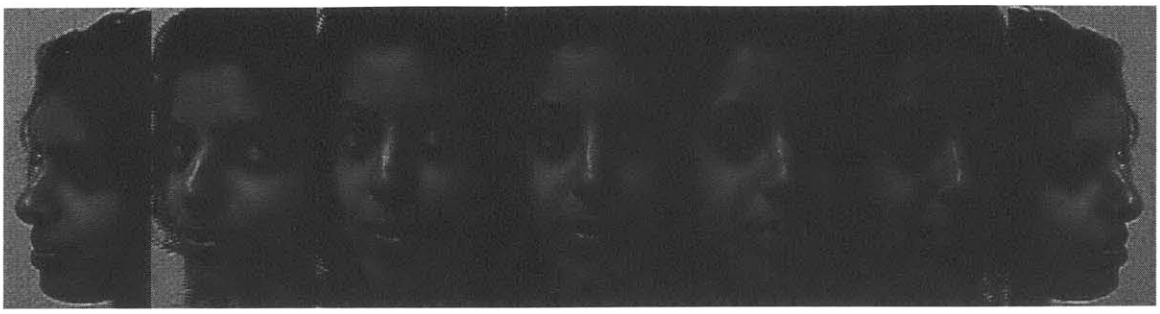


Figure 3: Sequence of 7 image of a neutral face rotating through 180 degrees.

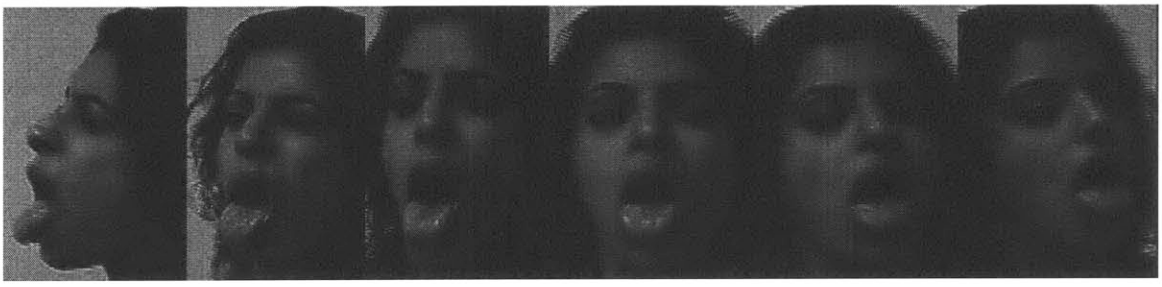


Figure 4: Sequence of 7 image of an exaggerated face rotating through 180 degrees.

Figure 5: Sequence of 7 image of an exaggerated, tilted face rotating through 180 degrees.

4.2.1 Disadvantages of video and alternative methods for performing the experiment

There were several unanticipated disadvantages of using a video camera that became apparent while viewing and de-compiling the movie. The first is that video cameras produce images of low resolution, which make feature extraction during the modeling phase more difficult. Repeating the experiment using a digital still camera would have produced better quality pictures because of the higher resolution available in the consumer product. However, using a still camera would require the person to pause at certain positions which would have taken more time, thus making it difficult to keep an expression consistent throughout a complete revolution. A high-speed video camera would have been the ideal method for this experiment, as it rapidly captures frames with high resolution and would give detail that is not possible with a traditional video camera.

The second was that the distances between the camera and the model varied with orientation, thus some pictures appeared closer and bigger than others. This could be corrected by moving the video camera around the subject. In our experiment, we edited the photos in Photoshop to produce images that were consistent.

The third method would be to use a 3D scanner. The advantage of 3D scanning is, although scanning a full rotation takes more time than videotaping, the overall process—which includes assembling the model using the associated polygon-editing software--is less time consuming. The details of the 3D scanning experiment are outlined in section 4.4.

4.3 Modeling faces in 3D using Maya

According to projection theory, the minimum requirement for modeling a 3D face in Maya[6] is using two orthogonal still images: one of a frontal view and another of a profile view. Using the Mpeg-4 standard as a reference for feature points (shown in Figure 1), Zhang *et al* use two orthogonal still images to form a wire frame for a 3D model of a face on Poser [11].

Using Maya, we employed a similar method for modeling the face. We chose seven frames as shown in Figures 3, 4 and 5, where the angle between consecutive frames is about 25 degrees and the sequence includes two profile images, a single frontal image and four other images at different orientations. The reason for choosing 7 frames instead of the minimum requirement of two is because it provides more information about the positions of the features during a revolution. In addition, the method that uses only two orthogonal images assumes that a face is symmetrical where, in reality, faces are not. Thus, using more images would produce a more effective clone of a real face.

After we uploaded a sequence of images into Maya, seven virtual cameras were set up to project an object—that would later become the face model—onto seven 2D screens, where each screen contained an image. This section will be explained further and will be made clearer in section 4.3.1, which discusses the importance of 3D projection geometry and why it is required in computer graphics.

---

[6] The same requirements apply to other 3D software.

4.3.1 Projection Geometry for Modeling in 3D

A 3D geometric description of a scene is the basis of recreating a virtual replica. This description is available from 2D projections, which give the necessary visual detail and geometric information.

On a day-to-day basis, we view objects on 2D display units such as television screens, computer monitors, books and photographs. These 2D displays are useful for representing objects in both 2D and 3D; on the other hand, 3D display units do not exist, which results to the complexity of displaying 3D models: a 3D object must be transformed onto 2D projection screen.

Every feature on an object has 3 parameters to describe its position geometrically, where the 3 parameters correspond to the *X, Y* and *Z* axes When a feature is projected onto a 2D screen, its *X* and *Y* coordinates are transformed onto the plane, and the Z coordinate represents the depth. In linear algebra, this transformation from a 3D to a 2D plane is defined in Equation 3.

$$Projection \sim R^3 \rightarrow R^2 \qquad \textbf{Equation 3}$$

In Maya, 7 virtual cameras are set up approximately 25 degrees apart in a semi-circle to define viewing directions and place the 2D snapshots of the model (that is to be built) in the correct view-ports. Figure 6 shows a simple illustration of this, using three cameras that project three images.

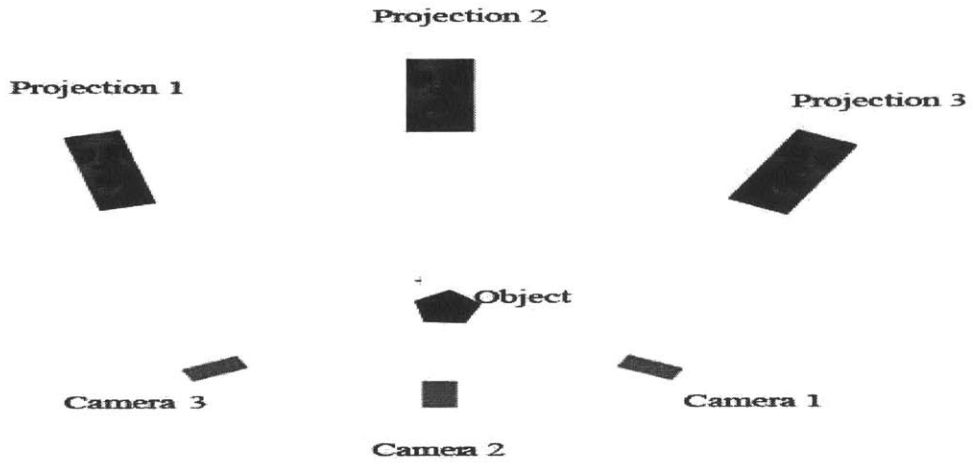


Figure 6: Illustration of how virtual cameras are set up in Maya.

Figure 7 is an illustration of perspective geometry, where a center of projection forms behind the projection plane (with respect to the camera and object positions). Using geometry of similar triangles, where *A(x,y,z)* is a 3D point on object, *A'(x',y',z')* is the projection point and *d* is the distance between the center of projection (COP) and A', the relationship between a point and its projection is defined as:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \rightarrow \begin{bmatrix} \frac{x}{z/d} \\ \frac{y}{z/d} \\ d \end{bmatrix}$$

Equation 4[7]

[7] Equation 4 assumes that the COP lies on the z-axis.

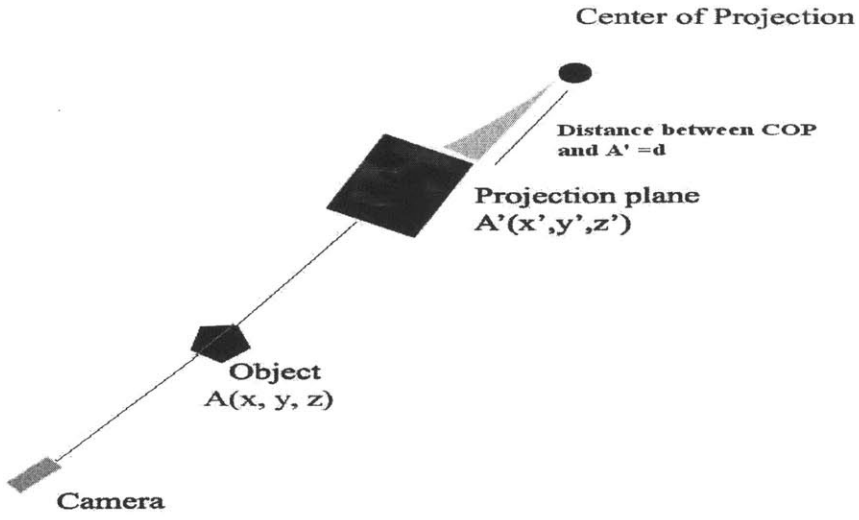


Figure 7: Illustration of projection of an object onto a plane.

Equation 4 shows how the z-axis remains constant and equal to *d* for all points of an object that are projected onto that plane, which demonstrates how projection transforms an object from a 3D space onto a 2D space when it is constrained by position of COP[8] [12].

The position of the same point on a different projection plane requires translation and rotation of the camera, which can be calculated using a translation matrix T and a Rotation matrix R, where T and R are 4x4 matrices [13].

$$[x' \quad y' \quad z' \quad 1] \times T \times R \qquad \textbf{Equation 5}$$

[8] This equation is only valid if the COP lies on the z-axis.

4.3.2 Constructing the wire frame

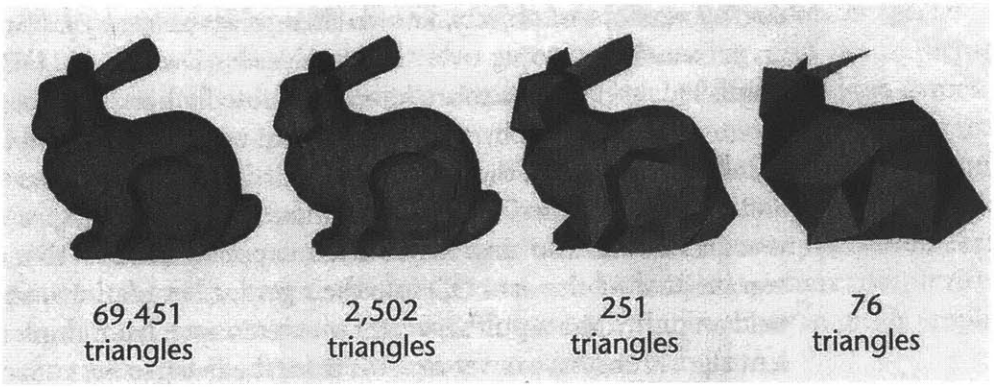


Figure 8: Level of detail of a 3D model relative to the number of polygons that are used[9]

A common method of reconstructing a 3D model from 2D images is by tracking image silhouettes, which outlines an object's extremities throughout consecutive frames. This method recovers the volume of the object. The geometry of an abject can then be found using algorithms that make use of the intersection of silhouettes [15].

The method we used to reconstruct the face from a video sequence on Maya also depended on the silhouettes from consecutive images. The Mpeg-4 standard for facial animation defines 84 significant points on a neutral face which mainly consist of features such as the nose, corners of the eyes, chin, jaw, mouth, ears etc. Figure 1 shows the 84 feature points defined by Mpeg-4. Since the still images we use are not of a neutral face, the feature points did not match exactly; nevertheless, we used this standard as a general reference for facial geometry.

---

8. Picture from Level of Detail for 3D Graphics [14]

3D surfaces are constructed using polygon meshes, where each mesh contains many vertices that hold all the geometrical data of the surface. Figure 8 shows four wire frames of the same rabbit, where the first consists of more polygons, thus containing more geometrical data. This results to a smoother, realistic looking surface that portrays more detail. We also used high-polygon meshes to build smooth, detailed models. Figure 10 shows our results.

The first step was to place an ellipse over the facial projection that corresponds to the image at --90 degrees, and to manually select vertices that defined key feature points. The vertices were then moved individually, deforming the ellipse until it took the shape of the facial silhouette. Since adjacent meshes are connected with non-rational, uniform B-splines (NURBS) which are piecewise continuous curves, the interpolation throughout the surfaces was smooth to the second derivative [16], and the object did not have unwanted sharp edges.

After the deformed mesh fitted the first projection, it was used on the next projection. Keeping track of the facial silhouettes such that they remained in the correct position, the mesh was modified again until it took the shape of the extremities on the second projection.

This process was repeated for all the images in the sequence. The end result was a mesh that traced the extremities from all seven projections, thus portraying the overall facial geometry. But since some features are obscure and do not change noticeably from one projection to another, they were not defined in vast detail during the mesh deformation process. To account for these "missing" features, Bezier curves were introduced to the model.

4.3.3 The importance of Bezier curves in animation

Achieving photorealism for special effects in movies, video games, art and other media applications is a central research area in computer graphics today. A requirement to achieve this realism when modeling a face in 3D is for it to behave in a natural and realistic way: certain features need to deform in the same manner as a real face. As mentioned in section 4.3.1, some features were not accounted for during the acquisition of extremities from projections since they were not distinctive. To modify these features, Bezier curves were introduced to the model.

The smoothness property associated with Bezier curves is a result of its similarity to NURBS; more specifically, Bezier curves are cubic polynomials which create a curve using four points: two that represent a starting point and an ending point, and the other two act as weights, which specify direction and degree of curvature. Because the equation that defines a Bezier curve is a cubic polynomial, it is smooth and continuous in the first and second derivative, meaning there are no discontinuities or changes in slope. Another similarity to NURBS is that NURBS result to a smooth object overall (see section 4.3.1), and Bezier curves cause smooth transitions during changes in facial expressions.

In animation, in order for facial movement to be believable, they have to mimic a reality. For example if a person speaks, it is not only the lips that move but also the facial muscles around the mouth. This type of deformation can be depicted by creating Bezier curves that map facial muscles. Therefore, studying the facial muscle anatomy is a valuable step prior to introducing the Bezier curves to the model, as it will allowed us to add the curves more accurately.

Figure 9: An illustration of the facial muscle anatomy

We used two methods side by side to investigate feature and muscle changes associated with changing expressions. We studied the shape of facial muscles, and mapped them onto various de-compiled video frames that were taken earlier in the experiment. This helped us gain knowledge about how muscle contractions effect facial features and which muscles are responsible for creating a certain expression. Furthermore, it helped us infer how to move Bezier curves to vary expressions.

Figure 9 shows as simple illustration of the muscle anatomy of a neutral face. The muscles can be described as elliptical around the eyes and mouth, angular around the cheekbones, between the lips and between the eyes, vertical on the forehead and the chin area and horizontal towards the ears [17].

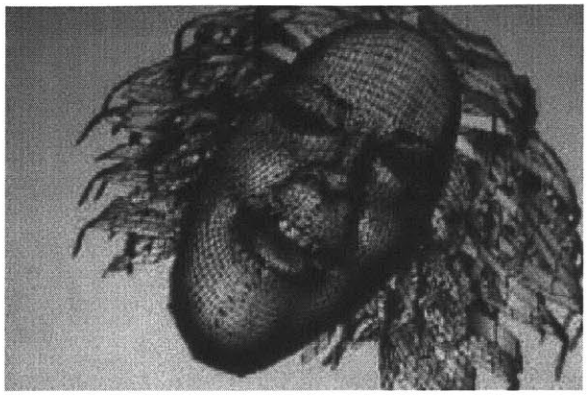


Figure 10a: Wire frame of a smiling face

Figure 10b: Wire frame of an exaggerated facial expression

Using a similar analysis, muscle lines were drawn around the extremities of two faces, a smiling one and one with an exaggerated expression, concentrating mostly on the mouth and jaw areas which tend to deform more. Figure 10 shows the musculature. Examining the differences between Figure 10a and 10b, we can see that in the latter the ellipse around the lips becomes tilted to some degree such that its major axis is no longer horizontal. More subtle differences are around the eyes, where the ellipses become stretched horizontally.

As shown in Figure 10, the wire frame consists of high-polygon meshes, which results to a detailed smooth model.

4.3.4 The importance of Bezier curves for Face Recognition

Section 4.3.2 stressed on the importance of achieving photorealism for animation but similar methods could also be of use for computer vision. The main disadvantage is that the process is manually intensive, thus consumes more time than what it required for computer vision. Nevertheless, computer graphics has been previously used for face recognition. In their research on improving face recognition, Blanz *et al.* used a graphics software—Poser—to model a face, and then applied face recognition algorithms to serve their purpose [11].

Since expression recognition is a problem that is yet to be enhanced in computer vision, Bezier curves may be of use. The common method of recognizing a face is by categorizing it as one of the six common expressions--happiness, sadness, disgust, anger, surprise, and fear—depending on the geometry of the face and location of the features. But even when this method is used, the error rate is still higher for expressive faces and furthermore, an expression may not fall within the six common ones. Therefore, using computer graphics and more specifically, Bezier curves, would allow one to maneuver the "virtual muscles" to go from any expression to a neutral one, and thus increasing a machine's recognition ability.

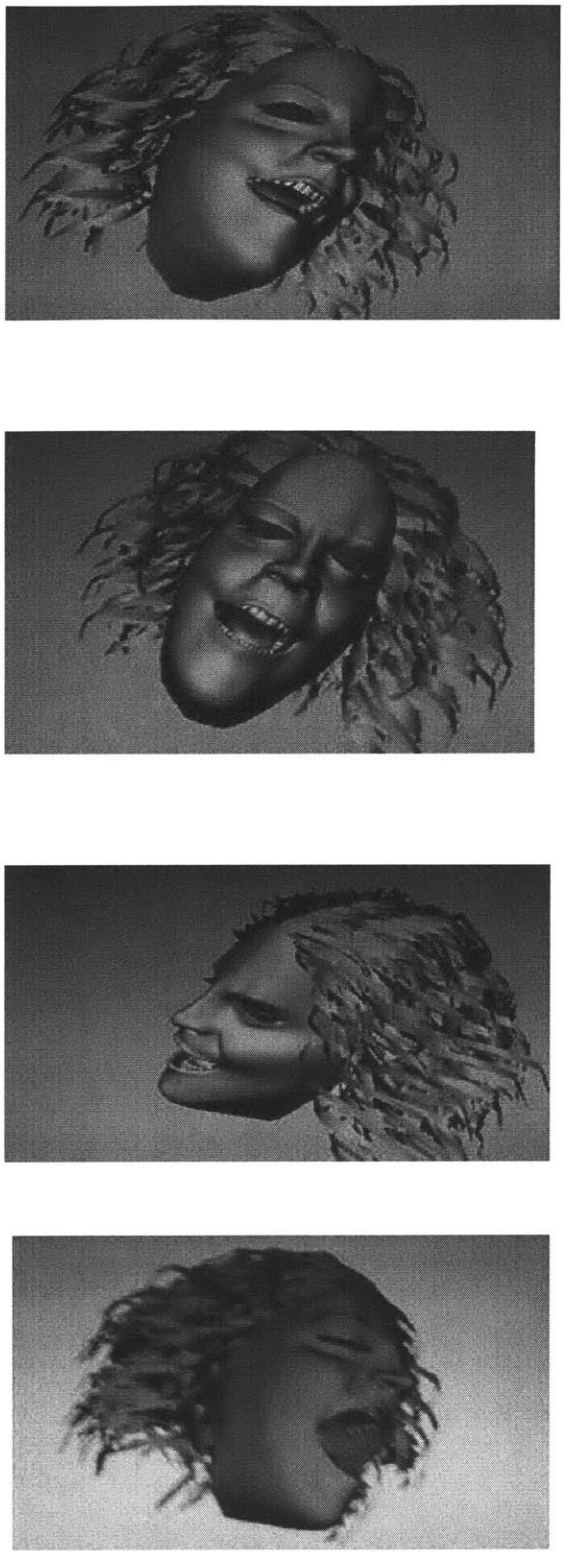


Figure 11: Expressive 3D Face models

4.4 Modeling 3D hair

Hair, or the lack of thereof, is an important feature of the human head, and can make a model more recognizable—if the aim is to replicate a person—and more realistic-looking.

Section 4.3.2 explained the procedure for animating a human face and the dependency of accurate modeling on the knowledge of facial geometry and facial muscle structure. Hence, the ability to model a face in a realistic manner has progressed significantly in computer graphics and computer vision; in contrast, modeling hair remains a challenge.

The ambiguity of hair geometry and the large number of hairs on a human scalp (approximately 100,000) are the leading reasons to why it is difficult to produce realistic looking and realistic behaving human hair. In the computer graphics industry, there exist many algorithms that model hair in real-time, but these methods have many limitations, including that they can only model straight hair realistically. The process of modeling hair to achieve believable-looking hair is a long and tedious process, where the same step for producing a single hair or a single strand is repeated continually until the right volume and right shape is achieved. In a recent movie—The Chronicles of Narnia--the lion's mane required six months of modeling[10].

In earlier animation, hair looked like a single patch and was modeled using a single primitive to reflect the overall geometry of hair: curly or straight, long or short. Advanced

[10] From talking to an animator that was involved in the process.

techniques use multiple patches that could be moved individually in order to create a more realistic animation.

In our model, we used cylindrical primitives, where each cylinder would represent a cluster of curly hair strands, as shown in Figure 11. Each primitive contains a skeleton such that the coil can take different shapes and be at different positions at different times. The advantages of this method are:

1. In a stationary model, we are able to attain hair volume and overall geometry.
2. For an animation, we can move individual patches in a way such that it resembles real hair motion during different activities like running or walking.

After modeling the hair, it is textured using 2D photographs of real hair that are applied to the patches.

Figures 10 show the complete models of the same face with different expressions and at different orientation. It can be seen, in the figure, that we were able to model fine detail, such as wrinkles on the forehead. In addition, due to bezier curves, the facial musculature varies during different expressions. The disadvantage of the model is that the hair did not look real because all the patches are identical in shape and size. In order to correct this, and achieve more realistic hair, one should use different shapes and sizes of primitives.

4.5 3D face using a scanner

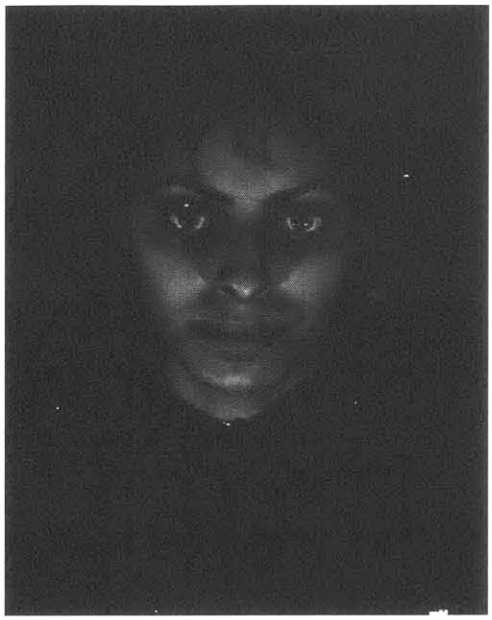


Figure 12a: 3D scan of a neutral face

Figure 12b: 3D scan of an expressive face

The main disadvantages of using 3D scanners is that they produce flawed results when there isn't sufficient illumination, which causes dark shadows on the face where data points cannot be registered. This results to a model that contains holes. In addition, as was discussed in the previous experiment, hair cannot be modeled using existing scanners mainly due to its ambiguous shape and small size.

This experiment was performed at Mitsubishi Electronic Research Laboratories (MERL) using a dome-shaped setup that contained 16 cameras to capture multiple 2D images at different angles, and 150 LED white lights, to supply sufficient illumination. During the experiment, a hairnet covered the hair so that it was away from the face.

Figures 12a and 12b show the results of the experiment, where 12a is an image of a neutral face and 12b is of an expressive face. Neither of the figures have visible holes, meaning that the system contained sufficient illumination. In Figure 12a, the model's only flaws are shadows around the nose. Whereas on the expressive face, the tongue is tilted and the teeth and eyes appear faulty and unrealistic.

In conclusion, this particular 3D scanner provided good scans of neutral faces, but caused errors in expressive ones. In comparison to modeling in Maya, the 3D scanner built a facial model in approximately 5 minutes, whereas modeling a face in Maya may take approximately 8 hours. On the other hand, Maya allows us to model any face regardless of the facial expression.

Chapter Five

Conclusion

The aim of this thesis was to explore methods in computer vision that would enhance the appearance of an animated model. The first phase of the experiment was researching methods in computer vision that would allow us to model a realistic three-dimensional face. The two popular methods for gathering data are using multiple images and using 3D scanners. Although both these methods have their drawbacks in the area of face recognition, neither has proven to be superior. We used both methods for the second phase: modeling expressive faces in 3D.

After selecting four sequences, each containing seven equally spaced images, we uploaded the pictures into Maya. Then, using the facial silhouettes from consecutive images, we modeled high-polygon, 3D faces. Since the silhouettes help in modeling extremities, such as the chin, the nose, the eyes, etc, we introduce Bezier curves in the shape of facial muscles around these extreme features, to represent less noticeable features. Placed in correct way, Bezier curves allow facial expression to vary from one to another in a realistic manner.

The second method involved using the 3D scanner at MERL, which contained 150 LED lights and 16 cameras. A common problem with 3D scanning is lack of sufficient light which causes holes in the 3D model, but because of the amount of light in the dome-shaped scanner, lighting was not an issue.

The advantage of the 3D scanner over modeling in Maya is that producing a model using a scanner took approximately five minutes whereas a model in Maya took about eight hours. The disadvantage is, although the scanner was able to replicate a neutral face, the expressive faces had some flaws such as misplacement of features. Maya allowed us to model any expression accurately. A problem that exists in both methods is the lack of ability to model hair realistically.

The next step will be to add an artistic, blur effect using various graphics programs to the models created in Maya. The 3D models with different expressions will be combined to form a single model, which will portray motion and changes in expression.

Bibliography

[1] R. Singh, M. Vatsa, and A. Noore. Performance enhancement of 2D face recognition via mosaicing. *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 63-68, October 2005.

[2] J.B. Colombe. A survey of recent developments in theoretical neuroscince and machine vision. *Proceedings of the IEEE conference on Applied Imagery Pattern Recogniton Workshop*, pages 205-213, October 2003.

[3] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* June 2005.

[4] V. Blanz, P. Grother, P.J. Phillips, T. Vetter. Face Recognition based on frontal views generated from non-frontal images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* June 2005.

[5] Chou, George Tao-Shun. *Large-scale 3D reconstruction: A triangulation based approach. MIT Ph.D Thesis*, June 2000.

[6] Y. Yacoob and L. Davis. Smiling faces are better for face recognition. *IEEE International Conference on Automatic Face and Gesture Recognition,* pages 52-57, May 2002.

[7] G. Passalis, I.A Kakadiaris, T. Theoharis, G. Toderici, and N. Murtuza. Evaluation of 3D Face Recognition in the presence of facial expressions: an Annotated Deformable Model approach. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* June 2005.

[8] W. Lee, M. Escher, G. Sannier, and N. Magnenat-Thalmann. MPEG-4 compatible faces from orthogonal photos. *IEEE Proceedings of Computer Animation,* May 1999.

[9] M. Ramachandran, S.K. Zhou, D. Jhalani, and R. Chellappa. A method for converting a smiling face to a neutral face with applications to face recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2005.

[10] A. Selinger, and R.C. Nelson. Appearance based object recognition using multiple views. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 905-911, December 2001.

[11] M. Zhang, L. Ma, X. Zeng, and Y. Wang. Image-based 3D face modeling. *IEEE International Conference on Computer Graphics, Imaging and Visualization*, July 2004.

[12] J. Foley, A. van Dam, S.K Feiner, and J.F Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, 1996.

[13] Gilbert Strang. *Introduction to Linear Algebra.* Wellesley-MIT press, 2003.

[14] D. Luebke, M. Reddy, J.D. Cohen, A. Varshney, B. Watson, and R. Huebner. *Level of detail for 3D graphics*. Morgan Kaufmann publishers, 2003.

[15] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 150-162, February 1994.

[16] Gilbert Strang. *Introduction to applied mathematics.* Wellesley-MIT press, 1986.

[17] K. Waters. A muscle model for animating three-dimensional facial expression. *ACM SIGGRAPH Conference Proceedings*, July 1987.