Technical Reports (CIS)   Department of Computer & Information Science

January 1994

Final Report to NSF of the Standards for Facial Animation Workshop

Catherine Pelachaud
University of Pennsylvania

Norman I. Badler
University of Pennsylvania

Marie-Luce Viaud
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/cis_reports

Final Report to NSF of the Standards for Facial Animation Workshop

Abstract

The human face is an important and complex communication channel. It is a very familiar and sensitive object of human perception. The facial animation field has increased greatly in the past few years as fast computer graphics workstations have made the modeling and real-time animation of hundreds of thousands of polygons affordable and almost commonplace. Many applications have been developed such as teleconferencing, surgery, information assistance systems, games, and entertainment. To solve these different problems, different approaches for both animation control and modeling have been developed.
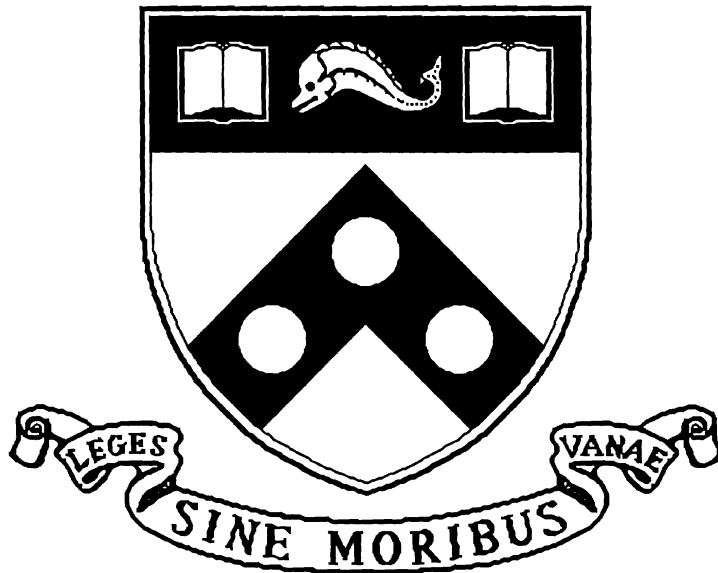
Comments

Final Report to NSF of the Standards for Facial Animation Workshop

Catherine Pelachaud
Norman I. Badler
Marie-Luce Viaud

October 1994

1 Acknowledgment

Contents

2 Motivation

Facial modeling and animation has long fascinated computer graphics researchers, not only for the ubiquity of faces in the real world, but also for the inherent problems in creating surface deformations expressive behaviors. Face animation is inherently a multi-disciplinary effort. Within computer graphics, applications of facial simulations have greatly increased as workstation performance permits real-time display of the hundreds of polygons necessary for minimal realism. Recent progress in facial animation now promises to provide useful and capable tools for virtual environments, entertainment, telecommunication, education, linguistics, psychology and medicine. Each of these fields has already applied various sorts of face models to aspects of research: as "talking heads", as computer-controllable experiment generators, and as plastic surgery mannequins.

Such a widespread demonstration of interest called for a workshop gathering together researchers in various disciplines such as computer graphics, linguistics and psychology. Noticing the multiplicity and diversity of the current research in this field, we felt the need to have a global view on the actual state-of-the-art and to define one or more "standards" in facial models. There has never been any similar forum with this purpose in this field. A relevant forum was held 1992 which was sponsored by the National Science Foundation: Planning Workshop on Facial Expression Understanding. The workshop reported on here looked at establishing a common database of facial expressions and at automatically extracting information from facial activity. This workshop can be viewed as a natural continuation of this first workshop; that is, given the analysis of facial actions, we propose to find common tools to synthesize the extracted data.

The goal of this meeting was to define scientifically defensible, computationally reasonable, and experimentally useful computational facial models as the basis for future research and development. Common facial models within as well as across discipline boundaries will accelerate applications, accessibility, interoperability, and reduce redundant developments, software costs, and animation control incompatibility.

The sponsors of this workshop are:

- National Science Foundation, Division of Information, Robotics, and Intelligent Systems.

- The Institute for Research in Cognitive Science at The University of Pennsylvania.

2.1 Introduction

The human face is an important and complex communication channel. It is a very familiar and sensitive object of human perception. The facial animation field has increased greatly in the past few years as fast computer graphics workstations have made the modeling and real-time animation of hundreds or thousands of polygons affordable and almost commonplace. Many applications have been developed such as teleconferencing, surgery, information assistance systems, games, and entertainment. To solve these different problems, different approaches for both animation control and modeling have been developed.

2.2 Modeling

Most facial modeling systems describe facial actions with either muscle notation or FACS (Facial Action Coding System) [37]. FACS is based on anatomical studies and denotes any visible movement.

- **Libraries of expression and interpolation:** In early models, modeling was done by digitizing sculptures of the face with various expressions (different lip shapes and expressions) and storing them in a library [41, 73]. Animation was performed by linear interpolation between given stored expressions. Such a method is really tedious and time consuming since it is not automatically adaptable to any other new model.

- **Parametric models:** Some other systems [105, 88, 108] were based on a set of parameters which affect not only the size and structure of the model (long nose, short forehead) but also its expressions (opening of the mouth, raising eyebrows). The separation between conformation parameters and expression parameters forces the independence between facial features and the production of an expression.

- **Physically-based models:** These systems describe the skin as an elastic spring mesh where unit actions are simulated by forces. The deformations are then performed by solving the dynamic equations [123, 144]. Muscle movement propagation is intrinsic to the model. Various layers of facial tissue are integrated [135]. It succeeds in producing subtle facial actions with realism.

- **Integration of additional features:** Texture mapping and hair modeling have been used to enhance realism of the model [151, 75, 2]. The consideration of wrinkles and of aging effects [138] adds much to the rendering of facial skin and expressions.

2.3 Animation Control

Animating every subtle facial action by hand is an extremely tedious task and requires the skill of a good animator. Automatic lip synchronization is included into animation systems by a hierarchical approach [66] or by adding speech parameters [60]. A correspondence between each speech unit and a basic lip shape is established. Some coarticulation effects are integrated [110, 28].

Three philosophical approaches have been taken to produce automatic facial animation: rules, analysis, and performance.

- **Rule-based approach**: Based on linguistics and psychological studies, a rule-based approach proposes automatic generation of facial animation. A set of rules describes the link between the intonation, emotion and facial expressions [110]. Multi-layer structures allows one to characterize various synchronous effects (lip movements, conversational signals, emotions and so on) [66].

- **Analysis-based approach:** This technique consists of extracting information from a live-video sequence and giving it as input to the animation system [134, 43]. Such information corresponds to muscle contractions or determination of FACS Action Units. Automatic extraction of facial parameters are difficult due to the subtlety and complexity of facial deformations and muscle correlations, but is a subject of much study in the computer vision community as reported by the Planning Workshop on Facial Expression Understanding.

- **Performance-based approach:** By tracking various points on a live actor's face and texture-mapping images onto an underlying polygon model, real-time facial animation synthesis can be achieved with little computational cost and no analysis [151] [[[plus recent system by SimGraphics]]].

2

2.4 Workshop Themes

The discussion at the Workshop had addressed the following questions:

- What is the vocabulary of signals (physical signals such as wrinkles, textures of the skin, elasticity of the skin, and so on, and expressive signals such as eyeblinks, smile, frown and so on) that characterize faces and their motions.

- Is there a minimal set of meaningful features for a good perception of faces?

- What are the requirements and specificities for different levels of representation, both geometric and functional of facial models?

- is there a more adequate notational system for the animation of computer graphics facial models?

- What are the minimal event times for facial animation or, equivalently, what are the bandwidth requirements?

- Should there be "standard" facial models? How many?

- How to validate a facial model and facial animation?

2.4.1 Significance

The Workshop will have great significance in providing an intense, focused, and broad-based discussion on the future course of facial models. Before the telecommunication thrusts of the '90s lead to incompatible and competing face animation systems, the principal developers should determine if common models can hasten widespread deployment and spur research in communication, medicine, education, virtual environments, and cognitive science.

Telecommunication : Considering the developments expected in the next decade in telecommunications and audio-video teleconferencing, the problem of sending compressed images around the world is a current field of research. Sending a minimal number of parameters controlling a synthetic model appears to be a promising approach. Therefore, a standard coding system would be highly desirable.

Medicine : Plastic, nerve, and muscle surgery, for example, aim for a precise simulation of skin behavior using a model integrating the various anatomical structures of the face. Facial reconstruction surgery is of great importance and would link the surface and muscle modeling community with the medical imaging profession.

Education and information assistance systems : There is an increasing need for user-friendly computer software. Speech synthesis is now becoming an important ergonomic component of some systems. Simple expressive faces are even appearing. Efforts beyond canned video are going to be needed.

Virtual environments : Recreation of particular (historical or popular) characters and personalities are coming to virtual reality! The entertainment and cinema industries are interested in synthetic actors for special effects and games.

Cognitive science : A tool that could analyze the significance of particular facial movements would help in cognitive, psychology, and linguistic studies. Conversely, some experiments require the precise control of facial expression. Even trained actors can have difficulty forming specific, non-standard, exactly repeatable facial expressions: having a tool offering such control is therefore very useful for experiment design.

2.5   Workshop Format

We proposed a meeting with a rather small number of invited speakers to provide the best opportunity for discussion. Each participant is chosen from computer animation, psychology and phonetics and have expressed great interest in and concern for the questions addressed by this workshop in their research and writing. The workshop was held on the 11th and 12th of November 1993 (Thursday and Friday). On the first day of the Workshop, each participant (or group), gave a 15 minute presentation, with 10 minutes for subsequent discussion. The second day was devoted entirely to discussion of the questions raised above and other relevant issues that the group had wished to pursue.

Workshop participants were asked to submit a 2-3 page position paper, which was distributed to the other participants prior to the Workshop.

3   Mosaic

A WEB server has been constructed at UCSC by Michael M. Cohen at
URL http://mambo.ucsc.edu
which includes information on work in the areas of facial animation, facial analysis, lipreading, and related topics. This information includes figures, mpeg movies, references, abstracts, papers and technical reports. A copy of the abstracts related to this workshop and this report can be found directly at : URL http://mambo.ucsc.edu/psl/fan.htlm
A WEB server at the University of Pennsylvania
URL http://www.cis.upenn.edu/ hms/home.html
contains a copy of this report.

4   Organization of the Report

This report starts with the motivation of calling for a workshop on standards for facial animation. The themes of the workshop is then presented as well as the list of questions the workshop tried to access. A description of format, schedule and a summary of the workshop follow. The list of participants and of the sponsors are given afterwards as well as the address of a WEB server.

The next sections relate to the three groups the workshop got divided into. Each of the group section give an overview of the field, and the different existing techniques. A sub-section on the problem of validation technique is addressed.

5 Modeling

Prem Kalra
Tsuneya Kurihara
Pamela Mason
Manjula Patel
Steven Pieper (group leader)
Demetri Terzopoulos (text supervisor)
Marie-Luce Viaud
Hussein Yahia

The purpose of this section is to (non-exhaustively) survey the facial modeling landscape. Part one defines facial modeling and lists several applications. The second part examines the object to be modeled—the face—and its attributes. Part three reviews techniques for facial modeling and deformation, while parts four and five discuss the role of data and control. The final part considers the validation of modeling results.

5.1 Modeling: Definition and Application

5.1.1 Definition

The complexity of the human face makes it a challenging subject for modelers. Facial modeling has been an active area of research in the computer graphics field for more than two decades. It benefits from and can contribute to the larger field of human body modeling. Facial modeling is also relevant in other fields, such as medicine and engineering. It is, in fact, a multidisciplinary effort.

A facial model is a mathematical abstraction that captures to some degree of accuracy the form and function of a face, whether human or otherwise, in a way that makes the model useful for specific applications. State-of-the-art facial models for computer animation attempt to represent the geometry, photometry, deformation, motion, etc., of the various organs and features associated with the face, as well as with the rest of the head and neck. These models rely on data from various sources (shape, color, elasticity, control, etc.). Typically, the models are designed to produce meaningful facial images.

5.1.2 Applications

[1]Typical models of the human face are relevant to a variety of applications, such as education, entertainment, medicine, telecommunications, etc. The amount of detail that the model captures is likely to vary from application to application.

Education In an educational environment, a major use of the face is in communicating ideas. For example, a model that captures the physics and anatomy of the human face may be used in teaching medical students about faces.

---

[1]M. Patel

Entertainment   The use of faces for entertainment often requires the elicitation of empathy and human emotion towards computer generated characters. The synthesis of facial expressions is important in this context.

Medicine   Preoperative simulation of corrective plastic surgery and dental treatment are of great interest to both practitioners and patients alike. Such applications demand precise models of particular individuals based on the bone and soft tissue of the head. A computerized system, which incorporates an anatomically complete model of the head and face, would provide surgeons with the capability to plan, and even rehearse, complex operations without undertaking costly and potentially dangerous exploratory surgery.

Narration   Speech is an integral component of human communication. A face model which incorporates speech synthesis capabilities could prove to be useful for the deaf and hard-of-hearing.

Telecommunication   Researchers are developing facial models for use in videophones (such as portable videophones) that must transmit facial images over low-bandwidth channels. A photorealistic model of the speaker is captured and transmitted to the receiving station where it is reconstructed at low bit-rates to produce a realistic animated image of the speaker's face.

Criminology   Recognition and identification of faces is an important aspect in criminal investigations. Here, representing the appearance of a wide variety of faces is particularly important.

Psychology   Many psychologists make use of computer generated faces to study human facial recognition and nonverbal communication. In addition, behavioral scientists may find computer generated faces a useful means of studying disabilities.

Forensic Medicine and Anthropology   Reconstruction of realistic faces from skeletal remains is of immense interest in forensic medicine and archaeology. Facial reconstruction can be employed to assist in identifying a victim from only a few clues. A computer-based system would require a complete model of the face in order to mimic the manual process.

Advertising   A major objective of the use of the face in advertising is to give the audience an unambiguous message. This requires accurate modeling of facial behaviors.

5.2   Properties and attributes of faces

5.2.1   Physical structures

[2]The face is a complex biological structure. The group found it interesting to attempt to create an exhaustive list of the physical structures of the face and their role in producing facial expressions.

The overall shape of the face is determined by the underlying bone shapes of the skull and the mandible (jaw bone). The bones are generally considered to be rigid in most applications of facial modeling, however it is obvious that changes in shape must be accounted for in any application concerning modeling of children or of the growth process. From a physical point of view, it is also

---

[2]S. Pieper

commonly noted in the medical community that soft tissue always shapes hard tissue – that is to say that if bone is compressed by muscle actions, the bone will eventually be reshaped in response.

The medical term "joint" refers to any region where two distinct bones come together. Several bone masses make up the skull, but by adulthood they have fused together to the extent that the jaw is the only feature of the face which fits our common sense definition of a joint as seen in other parts of the body. The jaw is referred to as the temporo-mandibular joint (TMJ). To a first approximation, the TMJ can be treated as a hinge joint. However, in practice it is important that the muscles control the lower jaw in all six degrees of freedom (this is particularly useful for producing grinding actions in chewing).

Several layers of soft tissue cover the bones of the face. Although the tissues can be categorized by function and material content, in vivo the difference between layers of tissue is less distinct (in any given volume of tissue, there may be muscle fibers interspersed with the collagen network of the dermis).

The muscles of facial expression tend to be of the flat, diffuse variety—more like the smooth muscles of the gut than the cylindrical muscles used for locomotion and manipulation in the arms and legs. Whereas the cylindrical muscles have well defined origin and insertion points, the muscles of facial expression have broad attachment areas integrated in the tissue. There may be several layers of muscle fibers connected to the same part of the anatomy (for instance the levator labii and the risorius muscles both insert at the corner of the mouth and are involved in raising it, but they differ in origin). Such muscles may or may not always be independently controllable.

The mechanical behavior, particularly the *Poisson effect* and the *elasticity*, of the skin and soft tissue is one of the primary determinants of the change of appearance with facial expressions. The Poisson effect describes the tendency of the material to preserve its volume when changing length. Since much of the mass in the soft tissue is water, the soft tissue is nearly incompressible. Thus when muscles cause a contraction along one axis, the face must bulge along another; since the underlying hard tissue forms a firm foundation, facial actions almost always cause the skin to bulge out from the face. This change in the surface becomes visible through changes in the silhouette edge of the face and through changes in the surface shading of the face. The other major mechanical effect, elasticity, is visible in expression through the displacement of features. When a muscle causes a movement at a particular point of the face (say the corner of the lip is raised), the tissue in the surrounding area is displaced also. The amount of displacement of a particular point is determined by its distance from the point being moved, the elasticity of surrounding tissue, and the influence of boundary conditions (such as a rigid attachment to hard tissue). In general, the Poisson effect and the elasticity of the soft tissue (represented mathematically by Poisson's ratio (ν) and Young's modulus (E) will be different depending on the material being examined. They also may depend on the orientation of motion with respect to, for example, the underlying orientation of fibers of the tissue. Therefore, these values should be considered to be multiple valued functions of spatial location.

The detailed response of the facial soft tissue to muscle action is determined by the distribution of types of material and the orientation of the fibers. In the absence of physical trauma or surgery, these conditions are determined by growth and aging processes. Obviously, the general shape of the face and the locations of facial features are determined by the developmental process. For an individual, there will be natural areas where a crease in the skin occurs, such as at a naso-labial fold. These locations are characterized physically as areas where the fibrous structure in the tissue is preferentially aligned along the axis of the fold. Similar asymmetric alignments of fibers may arise over time due to the mechanical breakdown of the tissue: age lines and wrinkles. These

features of the face occur along lines that are repeatedly exercised during facial activities. The process of wrinkle formation is similar to the fatiguing process in metals and other materials. Scars are characterized by a denser fiber structure and asymmetric fiber alignment.

5.2.2 Primary Facial Features

[3]The following features were identified as relevant in modeling the human face. The relevance of these features comes from their role in facial conformation, movement, and communication.

Nose   Nose movement usually conveys an emotion of disgust. In addition, nostril movements are observed during deep respiration and inspiration. The size of the nose varies among people with different origins. Nose shape contributes significantly to identification.

Eyebrows   Eyebrow actions play a vital role both in verbal and non verbal communication. They are predominantly visible in emotions such as "surprise", "fear", and "anger". They may also be used to accentuate a word, or to emphasize a pause or a sequence of words.

Eyes   Eyes are a crucial source of expressive information. When looking at a picture of a person, people tend to devote the greatest attention to the eyes. The eye movement may reveal "interest", or "attention" of a person. Eye blinks may occur to keep the eyes wet, or to emphasize speech, or to show an emotional state—hesitation, nervousness etc. The shape, size, and color of the eyes provide cues in recognizing individuals. The modeling of eyes should include the eyeballs and eyelids and their actions.

Ears   A face without ears looks like a mask. Ears have an intricate structure and shape. Modeling the detailed shape of ears may not be necessary, depending on the application. However, the simplification of ear shape changes the appearance of a complete face. Ear movement is extremely rare in humans.

Mouth   The mouth is a highly articulate facial zone. Lips articulate elaborately during speech. Modeling of lip motions should be able to open the mouth, stretch the lips, protrude the lips etc., to produce the phonemes and basic emotional expressions. The form and shape of lips is generally different for men and women. In addition, they provide attributes to distinguish different individuals.

Teeth   Teeth define the structure of a face as much as do the other bones; however, teeth are visible. Teeth modeling is needed for aesthetic, identification and dental surgery.

Tongue and Vocal Tract   Tongue movement is explicit, particularly in the context of verbal communication, in the formation of phonemes such as "ll", "dd", etc. The motion of the tongue often becomes obscured by the mouth motion. However, incorporation of tongue movement has immense importance for precise simulation of speech. The vocal tract is an important anatomical structure for speech production. This is of concern to clinicians.

---

[3]Prem Kalra

Cheeks Cheek movement is visible in many emotional states. Generally, cheek movements supplement other movements which may include the mouth or lower part of the eyes. The zygomatic muscles generate cheek movements while extending the corners of the lips when smiling or laughing. Actions such as the puffing and sucking of cheeks may provide emphasis for certain emotions. They reveal characteristic movements during sucking or whistling.

Chin The movement of the chin is mainly associated with jaw motion. However, the chin is distinctively deformed to indicate "disgust" and "anger" with the lips tightened. The shape of chin also plays an important role when conforming facial models to individuals.

Neck The neck permits the movement of the entire head, such as nodding, turning, rolling etc. As the neck moves, it can change its width or it may elongate.

Hair To complete the modeling of a face it is essential to include hair. The color and style of head hair is often an indicator of gender, race, and individuality. Hair modeling and animation is an active subject of research with tremendous relevance to facial modeling. Facial hair, including eyebrows, eyelashes, mustaches, beards, and nose hairs, is also important.

Accessories When relating to specific individuals, it is important to model accessories worn on the face and head, such as glasses, makeup, hats and hairpieces, and jewelry. People tend to see such accessories as identification marks.

5.2.3 Secondary Details

[4]In addition to the above features, the following are additional details which can greatly enhance a facial model, depending on the particular and application.

Nerves and Vessels For plastic surgery and other medical applications it is important to know the location of nerves and vessels and their possible interference with other facial organs, such as the facial muscles.

Physical Trauma and Surgery A facial model should incorporate the effects of physical trauma and surgery, such as scars, particularly when intended for such applications.

Vascular Expression Vascular effects such as blushing or pallor are also important communicational signals and messages. The facial model should include simulated vascular structures for synthesizing these effects.

Tears and Sweats Certain physical and emotional conditions are associated with moisture on the skin. For example, a person in grief or under intense pain may cry, producing tears. Similarly, physical exertion may cause sweating at the brow and elsewhere.

---

[4]Prem Kalra

Illness and Fatigue   Illness and fatigue are further examples of physical conditioning effects, where muscular tension and blood circulation may differ from the normal state and cause a physiological and visible change on the face.

Gender and Race   The facial model should give adequate information regarding gender and race. As one observes distinct differences for structure, shape, form and color of a face for different genders and races, it is essential that the model is able to convey these differences.

Growth and Aging   Growth and aging processes influence facial shape and structure greatly. The structure and size of bones changes, the skin texture changes, the skin fattens or sags, wrinkles appear, and even muscular activities change in terms of intensity.

5.3   Technologies and Representations

5.3.1   Inheritance and Hierarchy for Facial Models

The importance of genericity in facial animation.   [5]Facial animation by computer is a complex subject involving advanced techniques in human-computer interaction, physical and geometrical simulation, computer graphics and also image recognition. A living face (human or animal) in motion is a complex animated structure in which fine details, rigid and deformable structures run together producing a variety of expressions. The choice of a specific facial model is dictated by the particular effects one expects to achieve during animation. See [72, 105, 103, 98, 138, 119, 81, 101, 89, 66, 124, 144, 133, 146, 134, 148, 80, 75, 108, 28]. As discussed above, all human faces possess the same physiological organization of bones, muscles and skin. People smile in the same way—contracting the big zygomatic and the lips corner elevator. Nevertheless, all faces are different.

The complexity of facial animation is due mainly to the fact that there is no unique model integrating all primary and secondary effects perceived on a face: each particular aspect requires specific modeling (facial animation, wrinkles, speech, etc.) and all aspects intersect each other in a complicated manner. This is why the design of highly structured facial animation systems in which all aspects of facial animation are clearly integrated is an important issue. High level tools allow the design of facial animation from a generic point of view; that is, independent of a given geometry. Since it is now possible to obtain several geometric data bases representing faces, the design of facial animation would be greatly simplified if one could utilize expressions libraries that could be applied to any face.

High level tools and genericity.   The biological structure of the face reveals various notions of genericity at different levels. While building the face, features always "order" in the same way: the nose is in between the two eyes, the lips are under the nose, and so on. In other words, the morphology varies while the base structure is kept unchanged. The skeleton defines the overall shape of the face, while the muscles smooth and deform that shape. Skin layers influence in a noticeable way the final shape: fat accumulation in the dermis creates cheeks, double chins, and other phenomena.

Also, while deforming the face, each muscle is a generic object. However, its shape and contractions are sightly different from one individual to another. This generates slightly different

---
[5]H. Yahia

behaviors as muscles act. Some people smile with the extremities of the lips higher than the middle point of the superior lips, while others don't, but we detect a smile in both cases.

Wrinkles also carry a notion of genericity: the skin bulges under muscle compression, and wrinkles depend on the nature of the skin and on muscular contractions.

The process of aging is a complex one. It depends not only on the structure of the initial face but also on many aspects of one's life, including climatic, psychological and other parameters. Moreover, some aspects of aging are linked to the face's structure: fat accumulation, hollow places or wrinkle directions are well known [17]. Hence some generic aspects of aging can be drawn.

The idea of genericity is to use these properties of "similarity" between faces to generate tools that allow an easy creation of new synthetic faces from existing ones. Several systems provide various tools acting on an initial mask to create new faces [80, 107, 138, 143]. In addition, the new structure may inherit some parts of the deformation tools used to create facial actions. As soon as action units of the system are respected, higher level controls (expression libraries or sequences of animation) become available to the new model.

The design of orthogonal functionalities for facial animation which can be inherited and interactively modified frees a facial animation process from excessive complexity. It also makes facial animation software easily amenable to the inclusion of new developments.

5.3.2 Representation Techniques

[6]The interaction between the various layers of the face generates the complexity of facial deformations. Therefore, it is difficult to isolate representation techniques from deformation techniques: each model is an association between a geometric representation and some deformation tools. In some cases, as with the finite element method, the geometry of the model is strongly linked to the deformation method.

On one hand, techniques depend on the desired application. For example, the requirements for medical applications may be drastically different from the requirements for animation. On the other hand, it is often desirable to get as complete a simulation as possible of the entire structure (bones, muscles, skin, and internal actions leading to deformation are important). In certain cases, the visual effect (deformation of the external layer) is all that matters and the issues may be computation time and manipulation tools.

Polygons - For rendering reasons, polygons are often used. Several deformation techniques are associated with polygons:

- **Keyframing techniques**: [41, 33] Libraries of expressions are generated and the control is achieved by interpolation between two expressions.

- **Paramatric deformations**: [105, 103, 72] Deformations are grouped under parameters meaningful to the animator. Methods for control are varied, from numbers to data gloves (Section 5.5 and Section 7) Impressive results are obtained by puppeteers.

- **Muscle-Based deformations**: [124, 144, 107] Parametric deformations may be organized according to the arrangement of muscles in the face. The polygonal representation may be deformed by a muscle model, often taking the form of superposition of deformation basis functions.

---

[6]M.L. Viaud

- **Interactive deformations**: [75] Around 100 control points are selected on the face to control facial expressions. The propagation of the deformation generated by the displacement of these control points is calculated by projecting the face onto a 2D parametric space using a cylindrical projection. Then the 2D space generated by the set of control points is triangulated (Delaunay triangulation) and each intermediate point is recalculated: its displacement is the weighted sum of the displacement of the 3 closest control points.

- **Free Form deformations**: [65] Each deformation is associated with a deformation box (parallelepiped). The points of the face included in the deformation box are expressed in the coordinate system of the box. The position of the points are updated when the coordinates of the box are displaced. [129]

- **Simplex Meshes**: [32] Simplex meshes are connected meshes that are topologically dual of triangulation. Elastic behavior is modeled by local stabilizing functionals, controlling the mean curvature through the simplex angle extracted at each vertex. Skull, skin and muscles are represented with simplex meshes, characterized with a constant vertex to vertex connectivity. Furthermore, cutting surface parts may be easily achieved due to the local nature of simplex meshes: muscles attached to both skin and skull adjust the shape to the reconstructed skull.

Splines - For very smooth surface, a variety of spline-based surface patch methods can be used.

- **B-splines**: [98] The face is modeled using B-splines. Deformations are performed by moving groups of controls points.

- **Cardinal splines & springs**: [138] A cardinal spline representation is coupled with a spring network. Muscle deformations are generated by applying forces to the spring network. For each rest state of the spring network, the spline surface is recalculated to create discontinuities and bulges: tangencies are computed to keep the arclength of the spline segment identical at rest and under compression.

- **Hierarchical splines & springs**: [143] The face is modeled using hierarchical splines [49]. Muscles are defined by forces the definition points of which belong to the surface: muscle definition follows any face transformation. Additional effects such as wrinkles are provided by behavior maps.

Implicit surfaces [96] - Implicit surfaces usually create very expressive animations. This technique has also been used to generate symbolic descriptions of an object (a face by example!) by fitting simple primitives to range data of the object. First a primitive is given, then an energy function which measures the difference between the range data and the model is minimized each time a new primitive is added.

Physics-based models [124, 133, 121, 80] - Physics-based modeling is the most promising approach to achieving realistic facial deformations. Physical models of the skin have been proposed which make use of unilayer and multilayer spring meshes and finite element meshes. The multilayer models may include volumetric properties such as skin volume preservation. Deformations are induced by forces and simulated by solving the dynamic equations of the physical system. Several methods have been developed to define forces, including embedded muscle models. Physics-based

models are also suitable for realistically modeling the dynamics of hair and the fluid dynamics of tears.

5.3.3 Rendering

[7]Rendering is the process of taking a geometric model, applying lighting, a selecting a point of view (camera position) to the scene, and then creating a 2D image (bitmap or rasterization), or "snapshot" of that model. Basic rendering algorithms include wireframe, polygon shading, ray tracing, and radiosity.

Image library   One way to create a facial model is through the use of an image library which consists of a source library of predetermined images. The FACE project [5] uses this technique to display and animate a face model on-screen, providing a visual representation of facial muscle activity monitored during surgery. It uses bitmaps of faces in neutral and extreme positions, and muscles that are overlayed on the faces that contract and relax based on muscle voltage activity.

Image warping and morphing   Image warping and morphing are useful in creating facial animations because the movements can be very smooth and have seamless transitions from one state, or position to another. Control points, such as muscle-based models, or underlying mesh models or images, such as texture maps may be warped and morphed. Warping is the distortion method which distorts and changes an image. Using one face as the starting face and another as the target face, one may change facial expression through the use of control lines on both faces. This is more commonly used for facial models than morphing. A good example of warping is [108]. Morphing is also known as the transition morphing method; one object is actually changed into another object. Morphing is a combination of warping, tweening and dissolving algorithms.

Shadowing techniques   Shadows provide the visual clues for the real world interaction between light and the object. The physical world is full of light and shadow; without it, models appear artificial. This could be the desired effect, however, so shadowing is a subjective choice based on the application. More computation time is required to add shadows, but the benefit is in realism. They can be achieved through ray-tracing, radiosity, using a shadow Z-buffer or shadow maps. Shadows are important in facial animation, for example, in the look of heavy, dark eyelids for realism.

Texturing/shading models (multi-spectral)   Mesh facial models (either polygonal or parametrically based) may be given realism or texture by means of surface mapping and shading. Shading can smooth a polygonal model. Various methods are available and they may be applied alone or in combination depending on the desired appearance of the model:

- Flat shading: the pixels in a polygon are all the same color with no variation. If the model is faceted, each facet will be distinguishable. Flat shading is useful only as a low-cost rendering method.

---

[7]P. Mason

- Gouraud (smooth): This is a shade-interpolating method of shading that will make the object appear smooth, instead of faceted. This method doesn't work well with highlights or local light sources and one can often still see polygonal edges on the object.

- Phong: This is a normal-interpolating method. The object appears very smooth. This method goes a step further than Gouraud. A new shade is computed for each point, point by point before it shades.

- Bump mapping: is another method for producing maps of rough or textured surfaces, but it does not have the edge or shadow accuracy of displacement mapping.

- Displacement mapping: is a method for distorting a surface to produce an embossed or debossed surface that produces geometry with accurate edges and shadows. The displacement map specifies how the surface is to be moved before being mapped.

- Reflection mapping: gives the illusion of reflection or a mirrored effect.

- Environment mapping: is a method by which the model's surface reflects the environment on its surface.

- Opacity mapping: involves using the grayscale of a 2D object to define an object's transparency or opacity.

- Transparency mapping: gives the illusion of transparency, like looking through glass. This is particularly useful for skin pallor, as it is semi-transparent. For example, [65] includes an emotion model which expresses emotion through the vascular system, such as paleness due to fear or blushing due to embarrassment.

- Texture mapping: is the process by which the bitmap is applied, or decaled, on to the geometric model. Textures may be applied as either 2D bitmaps or scans [148, 80]. Photographs may be applied to mesh models as maps also as seen in [75]. The mapped textures may also be shaded in accordance with the lighting and surface geometry.

5.4 Hooks to data

[8]For facial modeling and animation, we need the geometry of faces, rendering properties, and deformation of facial expression or speech. For the acquisition of facial geometry, we have two major approaches: 3D input and 2D input.

5.4.1 3D Input

The use of the 3D digitizer is the most direct way of obtaining the geometry of faces. The laser illuminated triangulation method [31] involves a laser and camera. With this method, 128,000 range and reflectance samples may be obtained in a few seconds. Cylindrical projection is used for the measurement of faces. Williams [151] created facial models from measured data, and animated it. 120,000 samples are typically too much for rendering and animation use, so they should be represented by a simpler model. Fitting the obtained samples to a generic facial model is efficient for the facial animation. Waters and Terzopoulos [148] proposed a physics-based technique to

---

[8]T. Kurihara

reduce these samples to coarser, non-uniform meshes (see also [80]). One disadvantage of the laser scanner is that the equipment is relatively expensive.

Another 3D digitizing method uses 3D trackers. With this method, meshes are drawn on a face and the 3D coordinates of vertices are digitized using an electro-magnetic 3D digitizer. This procedure is not automatic and therefore is time consuming. The advantage of the method is that the polygonal mesh is designed according to the topology of the face, and then optimized (few polygons for a good definition of the shape). "Tony de Peltrie" from the University of Montreal [41], Marilyn Monroe and Humphrey Bogart from Daniel Thalmann and Nadia Magnenat-Thalmann [87] were created with this method.

CT (Computer Tomography) and MRI (Magnetic Resonance Imaging) are usually used in the field of medicine. These methods can capture not only the facial surface, but also inner structure such as bones or muscles. These additional structures will be useful for more accurate facial modeling and animation, as well as medical applications such as a medical operation simulation.

Geometric modelers may be used for interactive facial design. Facial structure may be modeled using standard computer graphics techniques, except several parts such as hair. Arbitrary facial models (such as imaginary faces or faces of historical person) can be designed. However, it requires time and design skill because faces have very complex structures. Commercial geometric modelers have been used for the face and body design of the figures in "Little Death" [40]. [100] has used interactive deformation techniques such as the "ball and mouse" metaphor [79] for face and body design.

5.4.2 2D Input

Many techniques have been proposed to acquire 3D shape from 2D images. Photogrammetry of stereo images can be used to determine the shape of faces. Two images of an object are taken from different viewpoints. Corresponding points are found between two images, and the 3D coordinates of the points are determined by triangulation. Automatically finding corresponding points is a difficult problem so meshes may be drawn on the faces. The first 3D facial models were created using this technique [103, 105].

The 3D shape of faces can be determined from a single 2D image of it, if a regular structure (such as a line array or a square grid) is projected on it [150].

Another technique is that a generic facial model is prepared and transformed to a specific facial model that is consistent with photographs of an individual face [75]. The advantages are that no meshes need to be drawn on faces and that no special hardware is required. The drawback is that it is basically impossible to acquire the detailed structure.

5.4.3 Expression libraries

Facial animation can be generated using keyframing, parameterization, or physics-based facial modeling. To represent facial expression, an expression library is convenient. The Facial Action Coding System (FACS) [37] is the most common method to describe facial expressions. FACS describes the set of all possible basic actions (Action Units) performable on a human face. Some sample Action Units are Inner Brow Raiser, Lid Tightener, Lip Corner Depressor. A multi-layered approach has been proposed for efficient control of facial animation [66]

5.4.4 Speech/Sound

Synchronization between facial animation and speech is an important issue. Parke is the first researcher to demonstrate this synchronization [104]. Lewis has proposed automatic lip-sync techniques [81].

5.4.5 Rendering/Material Properties

Texture mapping is the most popular technique for face rendering. Texture information may be obtained with a laser scanner or from photographs. When photographs are used, multiple photographs may be combined for arbitrary viewing angle [151, 75]. Care must be taken because the photographs are influenced by lighting. A more accurate rendering model for skin has been proposed which considers reflection from layered surfaces due to subsurface scattering [59]

5.4.6 Real-Time Performance Data

Several techniques have been proposed for the acquisition of real-time/performance data.

Vision Several research papers have been written on the subject of obtaining facial expression data from images. The feature tracking approach, the pattern tracking approach, and the optical flow approaches are the three major ones. Williams [151] has proposed to acquire facial expressions in real-time using mirrors and dots on the faces. This is perhaps the most detailed performance based facial animation system to date. Terzopoulos and Waters [134] have proposed a method using the "snakes" algorithm with facial highlighting. Haibo Li, Pertti Roivainen and Robert Forchheimer [82] have developed a method which extracts affine motions from a facial image and maps it to a FACS-based model for teleconferencing applications. Essa and Pentland [44] use optical flow measurements within a control-theoretic approach to extract facial action parameters from images.

3D trackers Special hardware has been developed by SimGraphics Engineering Corporation to obtain facial expressions in real-time [53]. This either uses a device that is worn by the user to measure facial motion or infra-red markers which are observed by special cameras.

MIDI controllers MIDI is the standard for conveying musical information electronically. A MIDI keyboard is a convenient tool and it is effective for controlling facial expression in real-time. Kalra [65] used MIDI keyboards and a DataGlove to control facial animation.

5.5 Hooks to Control

[9]The nature of the model will to a certain extent dictate the format in which signals are passed from the control module to the model to generate animation. For certain types of control, it may also be crucial that the model generate response patterns which simulate sensory feedback (this would be required, for example, if one were to attempt a physically realistic simulation of chewing that took into account the texture of the food).

---

[9]S. Pieper

A very general and natural way to think of the relationship between model and control is to mimic the human nervous system. For muscle control, this would require an input variable for each motor neuron to each of the muscles. The value of the motor neuron variable at a given instant would describe how much stimulation that muscle would receive. Aside from the fact that this scheme leads to a very large number of control variables, there is also the issue that muscles respond to stimulation in a highly non-linear way (e.g. several motor neurons for a muscle fire in sequence to maintain a constant muscle force, and muscle fatigue leads to a variable force output for a given level of stimulation).

A higher-level control method that has been tried in practice is to abstract beyond the level of motor neurons and deal directly with the muscle forces. For several facial modeling systems, the control interface to the model is expressed as a time varying force value for each muscle in the model. This method may still provide too many control variables for easy animation, and a control package may provide mechanisms by which the animator may create macro behaviors to control several muscle forces at once. From the point of view of the model, it may be important to limit the actual class of muscle force input that is acceptable; muscles don't push and only pull up to certain physical limits. Also, there may be constraints in the nervous system that dictate which muscles must operate as groups and which may be independently articulated. Some of this grouping information is encoded in the FACS system, and thus parts of the FACS system may be considered part of the modeling domain as well as the control domain.

Even beyond the muscle control level, some implementations of a facial model may include very high level hooks for describing expression directly. These may take the form of very broad parameters to modify several regions of the face with a single variable (such as Parke's expression parameters) or they may even abstract the facial model to a state-based control (as in an image-based user interface agent that shows different pictures of a face for "listening", "thinking", or "sleeping" modes).

In human facial expressions, the element of time plays an important role in communication. This parameter should also be available to the control module for use in planning the sequence of commands sent to the model. Depending of the complexity of the facial model, commands may have delayed impact on the shape of the face (for example, the muscles in the model may have a maximum contraction speed). For certain applications, the exact timing of certain events may be a critical goal (e.g. to synchronize to speech or music). One possible solution to this problem is to provide hooks in the model through which the control module can query the expected result of a particular control signal without actually invoking the control. A second possible solution is to express the control signal in terms of the desired result and build intelligence into the model to determine the exact method by which the result will be achieved.

5.6 Validation

Any judgment as to the validity of a facial model must be made in the context of the desired application. (Clearly the concept of "accuracy" is difficult to apply to an abstract cartoon representation of a face.) Validation processes are divisible into two broad classes: user-based and quantitative. User-based validation refers to asking the question "does it look right?" and in particular asking that question in the context of the particular application domain. For many applications of facial animation, this is the only feasible method of validation. Unfortunately, user-based validation does not always provide details about what is and isn't right about the model, and thus does not indicate a clear direction for improving the model.

The quantitative approach to validation can be employed in cases where the phenomenon being modeled can be measured. For facial animation of real people, there are several sources of data through which this could be done.

- Anthropometry data provides a database of measurements of the human body over the spectrum of race, gender, and age. In particular, *cephalometrics* is the study of measurements of the face. Cephalometric data can be used to determine in an average sense if the proportions in the model are "natural" for a human.

- Extensiometers are tools for measuring the relationship between load and displacement in tissue. For a physics-based facial model, a corresponding experiment may be done in the computer in order to provide a direct measure of the accuracy of the model.

- Dimensional Studies of Movement, as described by Ekman, provide a basis for comparing the movement of muscle groups in the model to normal values observed in a range of human subjects.

- Physiological Studies provide data about the characteristics of human behaviors in a form that may be cross-checked with the behavior of the model. For instance, the force and speed profiles of muscle movements have been studied in detail, as have the acceleration characteristics of the eye during saccades and visual pursuit. Speech behaviors have been particularly well documented.

- Optical tests provide a basis for validating the rendering performance of a model. Photographs of human faces taken under various lighting conditions could be simulated in the computer model and used to evaluate the shading model used to render the skin.

6 Recognition/Data Mapping

Alan Goldschen
Alan Kaplan
Eric Petajan (Group Leader, Text supervisor)
Marilyn Panayi
David Roy
Agnes Saulnier

6.1 Facial Feature Acquisition And Processing Techniques

6.1.1 Introduction

The static and dynamic features of a face model are either created algorithmically, or are derived from human facial data. This section describes the current facial data acquisition techniques and processing algorithms used to derive face model parameters from the human facial data. We avoid the temptation to describe higher level recognition algorithms such as facial expression understanding, visual speech recognition (lipreading), or face recognition. These topics are comprehensively reviewed in the report from the 1992 NSF Workshop on Facial Expression Understanding.

Face model parameters include head position and orientation, eye orientations, eyelid closure, jaw position, facial surface shape and texture (hair, ears, wrinkles), and a variety of oral cavity parameters (lips, teeth, and tongue position). Virtually every available sensing modality has been used to acquire facial data including optical, acoustical, mechanical, electromylogram (EMG), tomography, x-ray and dissection. The sensing techniques cover a complete range of spatial and temporal resolution, and invasiveness. No method for facial data acquisition exists which does not compromise in one or more important dimensions. For example, laser range finders provide good 3 dimensional surface maps but only of static objects.

The following description of data acquisition techniques is followed by an overview of processing techniques for general face parameters. Finally, methods for the derivation of visual speech parameters are described in a separate section due to the use of special techniques in a variety of other contexts.

6.1.2 Instrumentation for Facial Animation Data

A facial animation model needs either empirical data to feed the model directly or data which the model uses as the basis for a calculation. A texture map of skin and facial features is an example of the former; a model of the skull and attached muscles is an example of the later.

This short survey is concerned with the methods and apparatus used to gather the data for all such models. Some of these are used to drive animations of other parts of the body as well.

We divide the discussion into gathering photographic or videographic information, range and position finding, and medical imaging techniques that provide information about bones or muscle. In a few cases manufactures specifications are cited. This is not meant to be an endorsement of a specific product. Rather, it is intended to give the reader an idea of the state of currently available equipment.

The ideal system for gathering facial information would have the following characteristics:

- High resolution

19

- Three dimensional

- Real time operation

- Fully automated

- Require no special lighting

- Able to track a moving user in a large space automatically

- Not require that the user wear anything special or be in any way encumbered (non-invasive)

- Automatically register common reference points

- Inexpensive

- Good color accuracy.

6.1.3 Imaging Techniques

Photographic or videographic data are either used by themselves for facial animation or, more frequently, combined with other model data such as when a face is texture mapped onto a head shaped surface. These techniques may be two dimensional or three dimensional, with stereo views and processing.

Ultimately the photographic or video image must be reduced to a digital form. However, for some purposes a standard video image may lack sufficient detail.

Even with a single photographic image there are problems of registering of the image with appropriate points on the model. These are most often solved by hand work. Some 3-D scanners take video images simultaneous with rangefinding and produce a 3-D model with appropriate pixels associated with each portion of the surface. The Cyberware scanners are examples of this type. If multiple images of the same face are photographed and one wishes to use these in an animation the faces must be carefully registered one with the other and be of the same size else artificial movements will be introduced or the entire head could appear to pulsate. The manually located fixed point for a series of facial expressions or for a series of visemes (visual speech elements like phoneme) is normally the center of the bridge of the nose. This is the point that should be in registration to avoid artifacts. Nostril tracking has been used successfully for automatic registration and tracking of the mouth area [117, 118, 15, 116, 51, 56]. Even if the photographs taken are lighted consistently, variations in processing and more importantly in the flat bed scanners will cause variations in intensity and color among the scanned images. These must be matched for color and brightness or they will cause even more annoying artifacts if several facial images are used. Also the images must be lighted consistently, particularly if the subject moves. One can introduce shadows that appear to jump with a change of facial expression.

When stereo imaging is used special lighting is often employed. This will be discussed in Section 6.1.4. Special lighting techniques are applied in computer lipreading systems in order to make the oral cavity have sufficient light and shadow. For these applications the camera to subject position has been fixed with a harness [15, 116, 51, 56].

Another photographic technique used in producing animations (rather than data gathering) is rotoscoping. A rotoscope is basically a combination of a movie projector and a movie camera. An image is projected on an animation table which then guides the production of a modified drawing that is subsequently filmed and is in registration with the projected image. This is a very old technique used first in production of "Betty Boop" cartoons.

6.1.4 Rangefinding and Position Sensing

In this context, rangefinding is measuring the relative position of a series of points of a static surface from the rangefinder; position sensing gives a rapid time sequence of relative x,y,z coordinate information and additionally gives relative orientation.

A number of physical principles are exploited in doing rangefinding and position sensing. These include optical sensing and triangulation as done in rangefinder cameras, and contrast maximization of an autofocused image, as is done in a modern autofocus single lens reflex camera. Also ultrasound is used for rangefinding in the manner of traditional SONAR. Magnetic field generation and sensing is exploited to produce position sensors. Both AC and DC field systems exist. Another type of optical rangefinder uses a laser and diffraction grating to shine a pattern on the object in question. Looking at the pattern's size and shape gives range information. Mechanical position sensors in which the user is connected to a kind of large extensible mechanical arm are also used.

Before discussing a few of the designs in a little more detail, we want to mention some considerations that are common to all such systems. Many of these concern whole body animation, but are all to one degree or another applicable to facial animation. A few definitions are also forthcoming. (A good review article on the subject of position sensing is [92].)

Before deciding on a position sensing or rangefinding system one needs to consider several things. The accuracy, resolution and registration are all important. Resolution is the smallest measurable change. Accuracy is the range in which a position is correct. Registration is the correspondence between reported and actual position and orientation. Accuracy, if it is relative to a previous position and orientation, does not imply registration since errors may accumulate. Other considerations are update rate - how often a measurement is taken and responsiveness - or how much lag there is between a movement and its reporting. There are also worries of how robust systems are in a real world environment. For example, magnetic trackers may be affected by metal or stray magnetic fields. The range of operation is also a consideration, though more for full body animation than for facial animation.

Systems may be classified into two basic types, orthogonal to the technology used in many cases. These are called inside-out, or outside-in. In an outside-in system the source is attached to the moving object and the sensors are fixed. An inside-out system is just the opposite. An inside-out system can be made to have better resolution by using lots of sources so that the sensors are in close proximity no matter where the object moves. Some systems do not have either sources or sensors on the moving object and thus do not fall into the inside-out or outside-in categories. These include various laser ranging systems as well as optical rangefinders that work by using multiple images and finding common points (and solving a simple trigonometry problem) to find range and the contrast maximizing rangefinders.

Direct optical rangefinding systems must reliably find common physical points from the two separate cameras. This is usually done using "structured light" in which a grid or a dot pattern of one or more colors is projected onto the face in order to make the problem of common point identification much simpler. A full pattern recognition on the face to find corresponding points with ordinary lighting is much more computationally difficult.

The laser rangefinding system made by Cyberware Inc. projects a vertical line on the face and uses this line for rangefinding. A separate camera is used to get color information. Either the object is rotated under computer control or the scanner moves around a fixed object in order to get complete three-dimensional information. A complete scan takes less than 30 seconds. We don't know of any facial animation application that uses contrast maximization, but this should be fairly

simple to implement. It may be slow in that the scanner has to move in both x and y directions unlike the laser rangefinders.

Other optical systems and ultrasonic systems use multiple sources and sensors. The relationship among the sources and among the sensors is known. Triangulation allows one to calculate the 6 degrees of freedom required (x,y,z and pitch, roll and yaw angles). Some acoustic position trackers [86] in addition to the time of flight information used to triangulate also measure phase of the acoustic signal. However, acoustic systems are sometimes sensitive to environmental noise.

Magnetic systems generate 3 perpendicular fields. AC fields, such as used by the Polhemus Navigation Sciences [125] are the most commonly used. Three mutually perpendicular coils generate the fields. The induced currents are measured in sensor coils and the nutating fields allow the 6 degrees of freedom to be calculated. Eddy currents from metal in the field can be a problem with AC magnetic systems.

Ascension Technology [3] makes a series of DC systems called Bird, Big Bird and Flock of Birds. The DC system (pulsed DC) avoids the problems of eddy currents. DC systems are sensitive to ferrous metals, however and they must carefully subtract off the earth's magnetic field to operate successfully. (Fields produced by these systems are of about the same value as the earth's field.)

6.1.5 Medical Imaging Technology

Many facial animation systems rely on medical data for input to the model. Such information as the shape of the underlying bone structure, the musculature, and a model of the skin are often used in these calculations. Much of this information is available in anatomy books, but if we wish to model a particular individual medical imaging technologies can help.

Determination of bone and other hard tissue 3 dimensional structure is easily accomplished by means of X-ray tomography. (Any use of X-rays or ingestion of nuclear materials, of course, involves some risk.) Soft tissue studies are usually done by means of magnetic resonance imaging (MRI) or positron emission tomography (PET) scans. PET scans typically measure a level of uptake of a material and are used for example to measure activity within the brain. This is likely of limited use for facial animation directly, but may be of interest to those studying facial understanding.

6.1.6 Direct Contact Sensing

Facial muscles are generally located quite near the surface. This means that their activity may be measured by means of surface EMG (electromylograms) in which the electrical activity is measured by means of electrodes on the skin instead of inserted needle probes.

A facial waldo [130] is a device for direct electro-mechanical sensing of facial movements. The advantage of this technique is that very little processing of the raw data is required. The disadvantages are low spatial resolution resulting in crude facial movements, extensive calibration procedures, and intrusiveness.

6.2 General Facial Data Processing Techniques

6.2.1 Geometric feature-based matching

Geometric feature-based matching uses a database with a model for each face (size and position of eyes, mouth, head outline, and relationships among these features). For each image all the inter-feature distances needed are calculated. The goal is to get a one-to-one correspondence between

the stimulus (face to be recognized) and the stored representation (face in the database). Features extracted by vertical gradients are useful in detecting the head top, eyes, nose base and mouth. Horizontal gradients are useful for detecting left and right boundaries of the face and nose. For each face a vector of features must be calculated and then recognition is performed with a nearest neighbor classifier.

There is a lot of work on trying to automate the extraction of facial features. One method used to combine the curves obtained by edge detectors is a multiresolution approach [30]. Knowledge of approximate positions of features at a given resolution is used to guide searches at a finer resolution.

6.2.2 3D features and principal component analysis

The detection of the 2D position of features is dependent upon the facial orientation, so for some applications it is better to detect 3D features. 3D positions can be extracted by stereoscopic measurements with the use of 2 images (front view and side view). After extraction, 3D positions are converted to 3D feature vectors defined on an individual head. The number of vectors can be reduced by defining a smaller number of principal components which are taken in combination to form the original vectors. To evaluate the similarities between 2 principal components, a membership function is introduced for every principal component. A face is recognized as the person whose sum of membership values is greatest.

6.2.3 Deformable templates

The method of template matching consists of directly comparing the appearance of a given image with a reference image by means of a suitable metric. To extract features of interest (eyes, mouth) from the data, a parameterized template can be used [153]. A prior knowledge of feature shape, such as the eye, is used in an energy minimizing scheme [48, 155, 154, 19]. These templates are specified by a set of parameters which enable a priori knowledge about the expected shape of a feature to guide the detection process. The deformable templates interact with the image in a dynamic manner. An energy function is defined. It contains terms attracting the template to salient features (intensity, peaks and valleys, edges). The parameters are obtained by a minimization of the energy function. This method is relatively insensitive to variations in scale, tilt and rotation of the head, and to lighting conditions.

6.2.4 Detection by Generalized Symmetry

The goal here is to develop an automatic method for locating facial features (eyes and mouth). The symmetry operator is a powerful tool for finding interesting points in arbitrary natural scenes without a priori knowledge of the world [127]. This method is robust even if the face isn't directly facing the camera, the position changes, or the background isn't uniform.

6.2.5 Frame Differencing

If the background is stationary, the face region can be isolated from the background by first capturing the background image and then subtracting it from each subsequent image. Regions of the difference image with high amplitude are assumed to be face regions.

6.2.6 Features vs. Templates

Most of the work on face processing has relied on either feature-based or template-based techniques [18]. Pentland *et al.* [136, 115] has addressed the face processing problem by using eigentemplates. A *view-based* multiple-observer eigenspace technique is used for face recognition under variable pose. In addition, a modular eigenspace description technique is used which incorporates salient features such as the eyes, nose and mouth, in an *eigenfeature* layer. This method also provides an automatic feature extraction mechanism.

6.2.7 Predictive Estimation

Some systems can track head motion by constrained motion. The first step still involves feature finding and then the conversion of 2D feature position measurements into 3D estimates of position and orientation of the head used as a prediction. An extended Kalman filter formulation [4] can be used to provide an optimal linear estimate for dynamic systems. This filter is computationally efficient (recursive) and based on physical dynamics. The accuracy appears to be as good as a Polhemus magnetic sensor system.

Other models of prediction could be used such as a Hidden Markov model.

6.2.8 Optical Flow

Optical flow in an image is produced by motion. This technique doesn't require preliminary feature detection. Motion rather than shape is computed [90, 152], a low-level processing of the pixel data is performed. It is a direct representation of facial action after proper smoothing is performed on the motion field. The optical flow based method is quite sensitive to noise and cannot capture rapid motion. However, the method of Essa and Pentland [44, 42], which employs a multi-grid optical flow coupled with a physical model of a face appears promising, and captures spatial and temporal variation of facial motion in sufficient detail.

6.2.9 Feature extraction

Some techniques used in face recognition can also be applied to facial direction detection (feature extraction or template matching). The direction can be obtain by 3D feature points and the distances between them [99].

The direction of the face can be defined by the normal of the plane defined by 3 points. The 2 first feature points are given by the light reflection of the cornea of the eyes and the third point is located at the tip of the nose.

6.2.10 Facial Markers

One easy way to find the head and specific deformations of the face is to put markers on several key points on the face and then to track them[99, 151, 108]. There are 2 classes of markers:

- passive markers (colored dots) which are not self-identifying but require calculation of relative position.

- active coded targets (temporal or color coding of targets) which can be easily identified.

6.2.11 Structured Light

Structured light illumination can be utilized to capture facial expressions using triangulation [34]. Colored dots of light on a regular grid can be identified rapidly. This information is used to identify the active muscle action groups. This method facilitates the location of wrinkles and other deformations of the skin.

6.2.12 Active contour model

Features can be used to analyze facial expressions. Deformable contour models track some predefined features of the face to recognize expressions [134]. They use deformable contour models (snakes) to track the position of the head and the non-rigid motion of the eyebrows, nasal furrows, and mouth and jaw in video images. A snake is an energy-minimizing spline guided by external constraint forces and guided by images that pull it toward features such as lines and edges [133, 13]. Snakes are used to track the movements of various objects, in particular, lip movements [67]. The dynamic facial muscle contractions are directly calculated from the snake state variables. The purpose of the analysis of facial image dynamics is to resynthesize facial expressions. The deformations obtained are realistic but for successful tracking the facial features must be highlighted by make-up.

6.2.13 Conclusions

There isn't one good method suitable for every application. Classical methods are based on extracting local 2D features in static images, but there are variations

- analysis of 3D features

- analysis of dynamic images (optical flow, Kalman)

For facial expressions there are not only changes in the eyes, mouth and contour of the face but also changes in facial skin texture. It is very difficult to capture the action of the facial skin because no salient contour appears on the image, so very few methods provide satisfactory data.

6.3 Visual Speech Feature Processing

Most face models used for face animation do not include the components necessary to properly articulate the oral cavity. This is due in part to a limited supply of oral cavity parameter sources. A face model which will be used by a human lipreader for speech understanding has more stringent requirements than one used only for natural looking speech. For example, proper modeling of the tongue is very important for lipreading but only moderately important for other applications. Determining the important visual parameters for speech understanding when combined with high to low quality acoustic speech is an active research area. In addition, the dynamics or temporal derivatives of visual speech parameters should be represented in the model [56].

The most common sources of visual speech parameters are text driven models of speech articulation. The front end of these models are very similar to traditional text-to-speech models except that visemes are generated in addition to phonemes. The output of an acoustic speech synthesizer needs to be synchronized with the visual speech which forces accurate simultaneous modeling of acoustic and visual speech articulation. This is also an active research area. Systems of this type are described in [27, 94, 95].

A few researchers have attempted to derive phonemes from acoustic speech and then map them to visemes as an aid to understanding telephonic speech for the hearing impaired. Unfortunately, phoneme classification errors are compounded in the mapping to visemes which is not one to one and a proper mapping has not been completely established. However, an advantage of this approach is the theoretical availability of emotion, prosody and non-phoneme sounds which can imply changes in facial expression. Natural language understanding of text is also being used to generate facial expressions [112].

Cartoon animators have been mapping acoustic speech to lip movements for over 70 years. In some cases, this was done directly from filmed speech using the rotoscope. Several systems have been developed within the last decade which automatically extract visual speech parameters from talking faces using a video camera and frame buffering. The simplest approach to extracting lip movements visually is by placing dots on the lips and tracking them from frame to frame for lipreading [16, 47, 46, 131]. The disadvantages of this approach are low resolution of lip parameters and invasiveness. Another approach which provides higher accuracy but is still invasive is to paint the lips to aid in grayscale thresholding [134, 57, 8, 91, 12]. Both the inner and outer lip contour are obtained with this method. The system described in [116, 51, 56] obtained a combination of inner lip contour and teeth/tongue reflection using a head mounted camera, fixed lighting, and nostril tracking to avoid face paint. This same system was used to obtain a pure inner lip contour by blacking out the teeth as described in [15]. The system described in [117, 118] was a precursor to the [15, 116] system but did not use a head-mounted camera and relied entirely on nostril tracking to locate the oral cavity in a sequence of facial images. Optical flow was used in [90] to measure mouth motion dynamics after manual definition of four rectangular mouth regions. Viseme recognition from manually windowed mouth images was used in [12].

The measurement of tongue position is very difficult. Only a coarse indication of forward tongue presence was available in [117, 118, 15, 116, 51, 56]. Detailed tongue information was obtained in [157] using electromagnetic receivers glued to the tongue and teeth.

7 Control

Ken Anjyo
Christian Benoit
Keith Waters (Group Leader, Text supervisor)
Justine Cassell
Micheal M. Cohen
Irfan Essa
John Hestenes
Pete Litwinowicz
Pamela Mason
Fred Parke
Catherine Pelachaud
Stephen Platt
Ken Shoemake
Mark Steedman
Carol Wang

7.1 Introduction and Overview

[10]The charter of this group was to investigate manipulation and control frameworks for facial animation. These frameworks should encompass a rich set of techniques applicable to a variety of face geometries such as those proposed by the modeling group. Ultimately, the goal would be to provide a complete and succinct set of guidelines for controlling facial animation.

Facial animation control is the mechanism whereby a model can be articulated. Selecting a particular facial control mechanism depends on the purpose of the animation; for example, a lip-readable facial model requires a high level of precise three-dimensional manipulation of the lips, mouth, teeth and tongue. However, facial image-coding requires only a relatively simple model with a focus on pixel manipulation. Obviously, it would be inappropriate to suggest a single strategy for both situations; there are, however, many overlapping techniques that apply to both applications, and this section explores such commonalities by example.

To date, facial animation control has focused principally on three sub-disciplines: (1) facial expression, (2) facial conformation, and (3) lip motion for speech synchronization. Facial expression modeling involves deforming a model of the face into a recognizable configuration, such as the six canonical expressions of happiness, anger, fear, surprise, disgust, and sadness. Conformation control specifies individual faces from the universe of possible face prototypes that can also be associated with morphological or long-term facial changes, such as growth or aging. Lip-synchronization concerns the coordination of the jaw, lip, and mouth parts with real or synthetic speech samples. While some control strategies have been developed within each sub-discipline, a coherent framework encompassing all three has not yet evolved.

The three sub-disciplines within facial animation typically use one of four distinctive control techniques: (1) 3D shape interpolation (2) ad hoc surface shape parameterization, (3) muscle-based, or abstract muscle models, and (3) physically based models. No one technique provides an optimum control strategy since each possesses particular advantages and disadvantages. Each technique is therefore briefly described in this section.

---

[10]K. Waters

Facial animation has also focused principally on synthesis. However, a great deal can be gained from processing and analyzing temporal sequences of real faces in motion. For example, it should be possible to control an animation from live camera input, or to derive facial control parameters from image sequences to validate a facial model's behavior. Despite a small body of research in this area, there is agreement that facial analysis will undoubtedly play an important role in facial animation control strategies.

Validation is an important part of any facial animation control strategy. Typically, the validity of a facial animation sequence is done by inspection; for example does it look right? This criteria is sufficient for semi-realistic, cartoon-based faces; as the realism of the synthetic facial models improves, we would like them to mimic reality as closely as possible. A more rigorous approach utilizes coding schemes, such as the Facial Action Coding System (FACS), to calibrate actions. The muscle-based abstraction provided by FACS maps well into a facial animation control strategy and has been used in a number of facial animation systems.

While FACS has provided a valuable foundation for facial animation, two shortcomings limit its functionality in facial animation: (1) muscular contractions timing is not available, and (2) a lack of detail around the mouth for speech synchronization. Ultimately, it should be possible to validate a muscle-based control strategy by quantifying facial motion from temporal images of real people's faces, parameterizing facial articulators, and validating the resulting model. In this form it would be possible to provide a closed-loop solution.

Finally, validating the accuracy of facial articulation plays an important role in facial animation. Without a quantitative measure of faces in motion, facial animation will remain subjective. A body of work within speech analysis sheds some interesting light on the problems associated with validating facial motion.

We organize this section of the report by first describing the basic techniques that investigators have used to date for facial animation. Next, we describe frameworks for animation control that investigators have used. Then we discuss temporal control issues, followed by a discussion of extensions to FACS. Finally, we discuss validation issues.

7.2 Basic Techniques

[11]This section describes four basic techniques used to date for facial animation control.

7.2.1 Shape Interpolation

Shape interpolation is a common and simple technique to control synthetic faces and stems from the early work of Parke [102]. The process operates as follows: A database of discrete facial postures is created from 3D digitizers, stereo-photogrammetry, or optical scanners [31, 137]. These facial data sets have the property of topological equivalence, such that a complete mapping can be found for each vertex in every facial posture. Once a complete set is derived, then inbetween interpolation of the [x,y,z] coordinates can be computed. Unfortunately, this technique is limited to the matrix of facial postures which prompted the development of parameterization schemes.

7.2.2 Key-Node Parameterization

The objective of facial parameterization is to present the user with a small set of control parameters for the face [103]. These parameters are hard-wired into a particular facial geometry, and the

---

[11]K. Waters

28

parameters are only loosely based on the dynamics of the face. For example, the expression on the brow from surprise involves the manipulation of five or six vertices of the facial geometry.

Alternatively, distances measured between non-anatomical points (such as lip internal height or width) can be used to predict 3-D parametric contour functions [58]. This is particularly useful in the case of the lips, which follow regular rules of deformation.

7.2.3 Muscle-Based

Muscle-based models, or abstract-muscle models, mimic at a simple level the action of three primary muscle groups of the face: (1) the linear, such as the zygomatic major, (2) the sphincter, such as the obicularis oculi, and (3) the sheet, such as the frontalis major [144, 145, 87, 106]. This approach has also been extended to B-spline surfaces [141, 143]. There are two distinct advantages for these models: they are independent of particular facial geometry and they map directly into muscle-based coding systems.

7.2.4 Physically Based

Physically based models attempt to model the shape and dynamic changes of the face by modeling the underlying properties of facial tissue and muscle action [122, 124, 120, 147, 133, 148, 149]. Most of these models are based on mass-spring meshes and spring lattices, with muscle actions approximated by a variety of force functions. These models are computationally expensive and difficult to control with force-based functions. Research has also been reported on modeling the physical properties of skin using a finite element approach [78, 121]

7.3 Framework for Control

[12]A significant component of facial animation concerns the development of controls supporting the widest possible range of facial expression and conformation. Ideally, it should be possible to build a single framework for facial animation control where the handles are intuitive and natural to use. Therefore this section describes some of the fundamental problems in achieving this goal.

Though not usually done, the development of facial animation could be viewed as two independent activities: the development of control schemes, and the development of control implementation techniques. The control schemes may be viewed as control parameterizations, in which case animation becomes the process of specifying and controlling parameter set values as functions over time.

However, the animator is usually not interested in the specifics of an algorithmic implementation. In fact, from the animator's point of view, three key issues are relevant: (1) What are the control parameters? (2) How are these parameters manipulated? (3) Are the control parameters adequate and appropriate? From the animation system implementor's point of view, the three key issues are (1) Which algorithms should be used to accomplish the facial animation,? (2) What control parameters should be accessible to the user? (3) How should these control parameters be provided?

To date, most of the work in facial animation has concentrated on specific implementation techniques; little work has been done establishing optimal control and interface functionality. Consequently, the control features provided by each specific implementation have been heavily influenced by the characteristics of the particular implementation rather than attempting to fulfill a

---

[12]F. Parke and K. Waters

well understood set of functionality and interface goals. This is characteristic of the field's infancy; therefore, questions concerning useful, optimal, and complete control parameterizations remain mostly unanswered.

It seems important to provide a hierarchy of control. One way to think about this is as a set of "nested" black boxes, where each black box has a set of external control handles implemented by an internal mechanism. The "black box" at one level is the "mechanism" for the next level of abstraction. Several interesting questions are: What are the different levels of control and what are the mappings between control levels?

There are two major categories of control: expression control, and conformation control. Expression control is of course concerned with changes of facial expression. Conformation control is used to specify a particular individual's face from the universe of possible faces. In the ideal case, these two control categories should be orthogonal. Conformation should be independent of expression, and expression independent of conformation.

7.4 Control Architecture Examples

[13]To date, a number of different control architectures have been proposed. In the context of this report we explore five different models. Within the working group there was a handfull of distinctive approaches, each with differing goals; therefore, this section is illustrated with five different techniques currently being used.

7.4.1 A Muscle Model

[14]In this example an abstract muscle model controls facial expressions on a facial geometry. The muscle model provides a direct linkage to muscle-based coding schemes, such as the Facial Action Coding Scheme, and is independent of a specific facial geometry [144].

The underlying process is as follows:

1. Muscles are described as belonging to one of three primary groups: Linear, sphincter, and sheet. Each group describes a specific geometric deformation function emulating muscle action on skin tissue.

2. The muscles control parameters are (1) location on bone and skin, (2) zones of influence, and (3) type of contraction profile (linear, non-linear). Each muscle functions independently.

3. Muscles are then labeled to correspond to Action Units (AU's) in FACS. Consequently, facial expressions can be created by orchestrating a sequence of muscle AU contractions.

4. Muscle control parameters are then scripted into a keyframe animation system. At each frame in the sequence, muscles are contracted that in turn deform the facial geometry. The resulting geometry is rendered in accordance with viewpoint, light source, and skin reflectance information.

---

[13]K. Waters
[14]K. Waters

7.4.2 Physically Based

[15]A hierarchical model of the face provides a natural and obvious set of control parameters for the face [133]. Conceptually, this approach decomposes into six levels of abstraction involving representations that exploit what we know about the psychology of human facial expressions, the anatomy of facial muscle structures, the histology and biomechanics of facial tissues, and the facial skeleton and kinematics.

1. EXPRESSION. At the highest level of abstraction, the face model executes expression commands. For instance, it can synthesize any of the six primary expressions within a given time interval and with specified degrees of emphasis.

2. CONTROL. A muscle-control process translates expression instructions into coordinated activation of muscle groups on the facial model.

3. MUSCLES. As in real faces, muscles comprise the basic acquisition mechanism of the model. Each muscle model consists of a bundle of muscle fibers. When fibers contract, they displace their points of attachment in the facial tissue or the jaw.

4. PHYSICS. The face model incorporates a physical approximation to human tissue, implemented as a lattice of point masses connected by nonlinear elastic springs. Large-scale synthetic tissue deformations are simulated numerically by continuously propagating through the lattice the stresses induced by activated muscle fibers.

5. GEOMETRY. The geometric representation of the facial model is a non-uniform mesh of polygonal elements whose size depends on the curvature of the neutral face. Muscle-induced synthetic tissue deformations distort the neutral geometry into an expressive geometry.

6. IMAGES. After each simulation time step, standard visualization techniques implement by dedicated graphics hardware render the deformed facial geometry in accordance with viewpoint, light source, and skin reflectance information to produce a continuous stream of facial images, the least abstract of the representations in the hierarchy.

7.4.3 Semantic (Speech/Linguistics) Driven

[16]Facial expression, head, and eye motion can be automatically driven from spoken input, thereby providing a high level programming interface for 3D facial animation. In this mode of operation a particular spoken utterance, with associated intonation and emotion, can be computed independently of the facial model. Once the computation is complete, a facial model can be articulated through the Action Units (AU) described by FACS notation system.

The process is as follows:

1. Phonemes are characterized by their degree of deformability. For each deformable segment, the algorithm looks for the nearby segment whose associated lip shapes influence it, using the look-ahead model for coarticulation [70]. The properties of muscle contractions are taken into account in two ways: (1) spatially, by adjusting the sequence of contracting muscles if antagonist movements (i.e., movements which show very different lip positions, like pucker movements versus lip extensions) succeed each other, and (2) temporally by noticing if

---

[15]K. Waters

[16]C. Pelachaud

31

a muscle has enough time to contract (respectively relax) before (respectively after) the surrounding lip shape. Both constraints act on the final computation of the lip shapes [111].

2. Starting from a functional group (lip shapes, conversational signal, punctuator, regulator or manipulator), algorithms can incorporate synchrony, and create coarticulation effects, emotional signals, and eye and head movements [113]. Rules generate automatically the facial actions corresponding to an input utterance. A conversational signal (movements occurring on accents, like raising of eyebrow) starts and ends with the accented word, while punctuator signals (such as smiling) coincide with pauses. Blinking is synchronized at the phoneme level. Head nods and shakes appear on accent and pause. The head of the speaker turns away from the listener at the beginning of a speaking turn and turns toward the listener at the end of a speaking turn to signal a change of turn.

3. Facial interaction between agents and synchronization of head and eye movements to the dialogue for each agent are accomplished using Parallel Transition Networks (PaT-Nets), which allow facial coordination rules to be encoded as simultaneously executing finite-state automata [24]. PaT-Nets can call for action in the simulation and make state transitions either conditionally or probabilistically. All face and eye movement behavior for an individual is encoded in a single PaT-Net. Each node of the PaT-Net corresponds to one gaze function. A PaT-Net instance is created to control each agent with appropriate parameters. Then as agents' PaT-Nets synchronize the agents with the dialogue and interact with the unfolding simulation they schedule activity that achieves a complex observed interaction behavior. Probabilities appropriate for each agent given the current role as listener or speaker are set for the PaT-Net before it executes. At each turn change, the probabilities affect actions accordingly.

7.4.4 Text-To-Visual/Auditory Speech

[17]This synthesis approach is a descendent of Parke's [103, 104, 105] parametrically controlled polygon topology synthesis technique, incorporating code developed by Pearce, Wyvill, Wyvill, and Hill [109] and Cohen and Massaro [27, 28] and is principally focused on the lower face.

The facial model includes a polygonal representation of a tongue, controlled by four parameters: tongue length, angle, width, and thickness. While the model is a considerable simplification compared to a real tongue, it does contribute a great deal of information to visual speech perception.

In addition to the tongue control parameters, a number of other new (relative to the earlier Parke models) parameters are used in speech control, including parameters to raise the lower lip, roll the lower lip, and translate the jaw forward and backward. Some parameters have been modified to have more global effects on the synthetic talker's face than in the original Parke model. For example, as the lips are protruded the cheeks pull inward somewhat. Another example is that raising the upper lip also raises some area of the face above.

An important improvement in the visual speech synthesis software has been the development of a new algorithm for articulator control which takes into account the phenomenon of coarticulation [28]. Coarticulation refers to changes in the articulation of a speech segment depending on preceding (backward coarticulation) and upcoming segments (forward coarticulation). An example of backward coarticulation is the difference in articulation of a final consonant in a word depending on the preceding vowel, e.g. boot vs beet. An example of forward coarticulation is the anticipatory lip rounding at the beginning of the word "stew". The substantial improvement of more recent

---

[17]M.M. Cohen

auditory speech synthesizers, such as MITtalk [1] and DECtalk, over the previous generation of synthesizers such as VOTRAX [140], is partly due to the inclusion of rules specifying the coarticulation among neighboring phonemes.

Our approach to the synthesis of coarticulated speech is based on the articulatory gesture model of Lofqvist [84]. A speech segment has dominance over the vocal articulators which increases and then decreases over time during articulation. Adjacent segments will have overlapping dominance functions which leads to a blending over time of the articulatory commands related to these segments. Given that articulation of a segment is implemented by several articulators, there is a dominance function for each articulator. The different articulatory dominance functions can differ in time offset, duration, and magnitude. Different time offsets, for example, between lip and glottal gestures could capture differences in voicing. The magnitude of each function can capture the relative importance of a characteristic for a segment. For example, a consonant could have a low dominance on lip rounding which would allow the intrusion of values of that characteristic from adjacent vowels. The variable and varying degree of dominance in this approach naturally captures the continuous nature of articulator positioning. This model, as implemented, provides the total guidance of the facial articulators for speech rather than simply modulating some other algorithm to correct for coarticulation. To instantiate this model it is necessary to select particular dominance and blending functions [28]. For example, when synthesizing the word "stew", the consonants /s/ and /t/ have very low dominance versus a strong and temporally wide dominance function for the vowel /u/. Because of the strong dominance of the vowel, its protrusion value spreads through the preceding /s/ and /t/.

For simultaneous visual-auditory speech synthesis from English text, this system uses a common higher level software to translate the text into the required segment, stress, and duration information to drive both the visual and auditory synthesis modules. To carry out this higher level analysis, the MITalk [1] software has been integrated with the facial synthesis software. In addition to providing phonemes, the syntactic and lexical boundary information from MITalk are used to control other facial behavior, such as eyebrow raising, blinking, and eye and head movements. Currently running on an SGI Crimson-VGX, the system can produce high-quality real-time simultaneous visual-auditory speech of up to about a minute length after a short pause.

7.4.5 Lip Contour Parameterization

[18]In this control scheme the lips of real human faces are analyzed to extract parameters that drive continuous functions that fit the shape of the lips [6]. Using video analysis it is then possible to synchronize a lip model with the natural voice of the speaker.

One of the unique features of this scheme is that the lip contour shapes, while highly deformable, follow some very regular rules. In fact, the coefficients for the continuous functions can easily be predicted from three anatomical parameters measured on the speaker's face: (1) the horizontal width and (2) the vertical height of the internal lip contour, and (3) the distance between a vertical profile reference and the lip contact protrusion.

The process is as follows:

1. The images of the lips of a real human face uttering coarticulated phonemes are first recorded and then geometrically analyzed.

---

[18]C. Benoit

2. From this analysis, a set of lip-jaw shapes, representing the "labial space" of a speaker, as well as relevant control parameters, are extracted [8].

3. The three above mentioned control parameters predict a set of continuous functions (polynomial and sinusoid) that best fit the frontal projection of the contours of the "viseme" set. The analysis is extended to 3D [58] where the equations of the lip contours in the coronal plane can be derived. The lip volume is created by linearly interpolating three intermediate contours in-between the frontal, internal, and external contours of the vermilion zone. For each of the five contours, a function approximates each horizontal projection. Two extra parameters, the distances between the vertical profile reference and (4) the lower and (5) the upper lip protrusions, are necessary to predict all the equations of the 3D model.

4. The model is animated and synchronized with the natural voice of the speaker whose lip gestures control the model by real-time video analysis [54].

7.4.6 Mix And Match

[19]One could argue that the choice of a particular control mechanism is inherently dependent on the purpose of the animation. Therefore, if facial animation control is partitioned into modules, it should be able to select and combine the techniques that satisfy the specified goals. This approach, for want of a better description, is called "mix and match". To illustrate this approach the control techniques are partitioned into a high and low levels of control.

Examples of the high level control techniques might include (1) a speech module taking in text and generating lip motion for a model, (2) a vision module capable of extracting facial motion from an actor and drive a computer-animated facial model, (3) a natural language module taking text and generating expressions, gaze direction, or head movements based on derived meaning, (4) a script-driven module taking commands such as "be mad", or "look right", instead of interactive input.

Examples of the lower-level control modules are more directly tied to the geometry itself; for instance, (1) a muscle-based system may be used to control facial deformations or (2) the face may be controlled by morphing software. The former animates from the inside out where an example command might be to pull on the upper lip muscle to open the mouth. The latter works from the outside in, where an example command might entail moving an upper lip line closer to the nose to achieve mouth opening.

Making a picture sequence is usually the last step, using various rendering techniques, which may be either 2D or 3D in nature. However, in some video tracking algorithms, the picture sequence is compared with the input video, and a feedback loop is used to further refine the tracked data. The following example scenario illustrates the "mix and match" approach:

1. An actor's facial performance is recorded onto video tape. Facial features are tracked from this performance using vision based software. The acquired motion is then mapped onto another character, via mapping software [108], and morphing software is used to animate the face. The tracking and morphing are done in 2D, to produce 2D facial animation.

2. Alternatively, an actor's performance is tracked. Lip movements are retrieved and FACS parameters are derived to then drive a computer animated FACS-based model. Note that in

---

[19]P. Litwinowicz

34

this case the specification of the movement is by facial features, but this specification is then mapped onto a FACS model to produce the deformation of the face.

In essence, one animator's low-level box may be another's high-level box, so the choice of what is low or high level is somewhat arbitrary. The lower levels are modules more closely tied to the specific type of model being used, whereas the higher levels can be model independent.

7.4.7 Vision/Performance Driven

[20]Facial analysis will undoubtedly play an important role in facial animation control strategies for two principal factors : (1) automatic control parameter extraction, and (2) validation of facial articulations. Deriving facial motion parameters from video images is an area of active research within the vision community (for more details see the NSF report: Facial Expression Understanding [38]). Therefore this section briefly identifies two key approaches used to date for facial animation:

1. Facial animation control has primarily focused on manually produced sequences by artists who carefully craft keyframes. This process is time-consuming and laborious. Therefore, automating facial parameter extraction is highly desirable and has resulted in a number of physical tracking approaches involving head-mounted devices attached to 6 DOF measuring devices [125]. The performance is then captured using multiple sensors and directly applied to facial parameters.

2. An alternative strategy is to derive parameters from real peoples' faces in motion using vision-based techniques. However, the current experience on facial motion tracking, and in particular expression tracking, is limited. Usually the work is restricted to frontal images of the face under suitable illumination. Furthermore, the ability to identify the position, orientation, and scale of the head and facial features in advance of tracking impedes progress.

7.5 Temporal Control Issues

[21]This section explores the problems associated with temporal issues in facial animation, since variations in this dimension alone can alter the semantics of the expression. For example, if an expression is slowly animated from neutral to happy, the resulting expression may be interpreted as insincerity rather than happiness. Consequently, script-driven systems, or natural language systems driven from textural input, have to integrate the temporal issues of motion.

It is important to differentiate three aspects of time for facial animation:

1. The timing of the onset, apex and offset for an expression.

2. Internal synchrony with speech, for example, intonation and emotion.

3. Inter-synchrony to take into account interpersonal interactions.

7.5.1 Timing of Facial Action

Facial movement can be described by three temporal parameters:

1. Onset duration: how long it takes to appear.

---

[20]K. Waters

[21]C. Pelachaud and P. Litwinowicz

35

2. Apex duration: how long it remains in the apex position.

3. Offset duration: how long it takes to disappear.

Unfortunately, when and how a movement appears and disappears, and how co-occurrent movements integrate with each other (coarticulation effects) are difficult to quantify. Once a change in facial expression is specified, the rate of change can determine the perceived meaning of the intended emotion. Duration of facial expression of emotion is often related to the intensity of the emotion. Facial expressions due to felt or unfelt emotion, or voluntary or involuntary facial expressions, differ in their timing parameters. A fake or polite smile might occur too early or too late; its onset or offset could be too fast or too slow. There is evidence to support that someone who is trying to convey an emotion methodically (trying to mask the real emotion, for instance) will have faster onset and offset than someone giving a spontaneous emotion [36]. Also, there is a report that studied some socially skilled and socially unskilled subjects and found that the socially unskilled people "may possess basic verbal communication skills but may fail to respond to the subtle social cues governing the timing of emotional displays" [55]. The sending and reading of the timing of expressions is important in determining the real emotion.

Few data and studies relate to the problem of computing the onset and offset of facial expression. The three parameters of a facial action (onset, apex and offset) can be obtained by human observers. However, these parameters typically lack precision for facial animation because facial action is a more complex pattern than a simple decomposition in three linear parameters. In the case of speech, facial expressions (not only lip movements but also eyebrow movements, smile and so on) and speech are coordinated by a synchronization phenomenon defining the apex of these facial expressions. For the lip motion, it is important to know when the target of the given lip shape should be reached. For example, is it on the pitch of the segment? Or at the beginning of the segment? Also when does the movement decay? Then the problem of coarticulation arises. The coarticulation of movements tries to achieve the required facial action goals in a given time; however, in some cases there is not enough time to accomplish an action. A few computer vision systems are being developed which are capable of extracting temporal characteristics of facial expressions. Yacoob *et al.* [152] tracks expression generation over time for recognition, while Essa *et al.* [42, 44] observes change in facial pattern and relates it to both muscle and FACS descriptions to extract coarticulation in facial expressions.

7.5.2 Facial Expressions, Speech and Emotion

Speech production and facial expression (including body movement) are linked with each other by a synchrony phenomenon [128, 29, 69, 20, 23]. Synchrony implies that changes occurring in speech as well as in body movements should appear at the same time. For example, when a word begins to be articulated, eye blink, hand movement, head turning, and brow raising can occur and finish at the end of the word. The effective presence of a movement and its intensity is related to the emotional state of the person. However, the time of appearance is linked to the synchrony phenomenon.

A person expresses his or her thoughts with words and expressions. For example, smiling, raising eyebrows, nose wrinkling co-occur within a verbal content. Some facial expressions accompany the flow of speech and are synchronized at the verbal level, punctuating accented phonemic segments and/or pauses. Other facial expressions emphasize what is being said, substituting for a word.

There are studies to support the notion that verbal statements can modify the interpretation of emotion in a facial expression. When congruent context statements are heard at the same time with a facial expression, people will tend to agree on what the facial expression's emotion represents. On the other hand, people will differ on implied emotional content in a facial expression when incongruent context statements are heard [97]. This is an extremely important observation for any facial animation that attempts to mix speech and facial expressions.

7.5.3 Dialogue Situation

During an interaction, the behavior of a person and the presence or absence of feedback by this person affect the behavior of the other participants. A conversation consists of the exchange of meaningful utterances and of behavior. For example the speaker punctuates, reinforces her/his speech by head nods, smiles and so on. Likewise the listener can interact by smiling, vocalizing or shifting her or his gaze to participate in the conversation. Facial expressions and gaze behavior help the interaction because they control the flow of speech and speaking-turn exchange. They depend heavily on the relationship between the participants, their personalities, emotions, attitudes, social identities and so on. All movements made consciously or unconsciously influence the other participants.

7.6 Validation

[22]This section describes concepts and techniques for validating control algorithms and for conveying linguistic and para-linguistic information by facial animation. By validation, we mean an evaluation of the degree to which synthetic faces mimic the behavior of real faces. To date, there is a considerable body of prior work that has been aimed at evaluating of the auditory communication channel. This section therefore describes some of this body of prior work.

Basically, there are two general approaches that can be employed:

1. Measurement based validation.

2. Perception based validation.

7.6.1 Measurement Based Validation

In the measurement based approach it is desirable to compare a natural signal with a synthetic signal to see how accurately it is possible to simulate the natural one. In auditory speech, for example, we might compare spectrograms to see how closely the patterns of formants (resonances) agree in frequency and temporal characteristics. Regarding suprasegmental information, we might compare the pitch contours of natural and synthetic sentences. For the auditory channel, many standard techniques for analysis of these characteristics are available. For example, FFT, cepstrum, and linear prediction (LPC) can be used to extract formant information. For the visual channel, the situation is less advanced and much less standardized.

7.6.2 Facial Measurement Techniques

A straightforward approach to comparing natural and synthetic faces would be to simply compare the images. However, except for some fairly limited circumstances (such as 2D picture warping),

---

[22]M.M. Cohen

the differences are too great to make this practical. Therefore, it is necessary to measure and compare features of the faces. A variety of measurement techniques have been developed which we will briefly discuss.

Based on speech production, classes of mouth/jaw patterns may be analyzed from video-taped corpora of natural speech. Several investigators have studied various shapes of speakers' faces ([50, 83, 21, 156] for American and French). Some authors have even related geometric measurements to visual confusions of lip shapes by subjects for American, Australian and French [62, 93, 25].

Early work on automatic analysis of visual speech information used a special purpose video image processor to characterize visual information [117, 118] and [15, 116]. Finn [47] made 2-dimensional hand measurements of facial points from video frames for use in automatic lipreading of VCV syllables. These measurements were transformed to five interpoint distances as input data for an optical recognition algorithm.

Matsuoka et al. [91] used black lipstick to make lip shape extraction easier. Similarly, Lallouache and Worley [77] and Lallouache [76] used blue lipstick, processed with a chromakeyer to preprocess the video signal. This enhanced signal was then automatically processed to extract articulatory parameters such as position and protrusion of upper and lower lip or lip opening area.

Stork et al. [131] gathered visual data consisting of the positions of ten reflective markers placed on the talker's face and sampled at 60 Hz. Measurements between the points were then used as inputs to a TDNN recognition system.

Tamura et al. [132] defined a spline model fitting scheme for the extraction of lip shape information. A similar approach was used by Kass, Witkin, and Terzopolous [68] and Terzopoulos and Waters [135] who used deformable contours known as "snakes" for tracking lip and facial feature contours. A related method of obtaining lip contours is described by Bregler and Konig [13], Bregler, Hild, Manke, and Waibel [12] and Bregler and Omohundro [14].

Mase and Pentland [114, 90] and Yacoob and Davis [152] measured motion rather than shape. They computed the optical flow of different regions of the face, and then extracted facial parameters in terms of the FACS model for recognition. The FACS model has also been used in automatic analysis, transmission, and reconstruction of facial images by Choi, Harashima, and Takebe [26] and Haibo Li, Pertti Roivainen, and Robert Forchheimer [82]. These investigators used a gradient method on all the pixels in the area of interest on the face to extract the FACS codes. In another approach using FACS coding Kaiser and Wehrle [64] used information from dots on the face as input to a neural net for pattern classification. Essa and Pentland [44] measured facial motion using optical flow and coupled this measurement with a physics-based facial model to extract detailed measurements both in space and time.

7.6.3 Validation Corpora

A preliminary question that needs to be addressed concerns the material on which the validations are based. There are a number of issues involved.

1. The 1st issue concerns the use of standardized, widely available corpora. We believe that, as with auditory communication, it is valuable to assess various systems with common standards. Benoit & Pols [9] give a detailed description of the various methods and techniques that are used for the evaluation of synthetic speech, including intelligibility, naturalness, pleasantness, and acceptability.

2. A 2nd issue concerns the selection of linguistic materials. For example, a number of different types of materials will be useful: letters, digits, syllables, words, pronounceable non-words, meaningful and non-meaningful sentences, and paragraph length samples. Phonetically constrained stimuli are especially valuable for analysis of recognition errors, for later improvements, in addition to an evaluation test. Additionally, the test materials should be produced with a variety of emotions.

3. A 3rd issue relates to the use of corpora (e.g., TIMIT) in which the segments are relatively balanced in frequency versus corpora (e.g., the Wall Street Journal) in which segment frequencies more closely reflect how often the segments occur in the language.

4. A 4th issue is talker variability. Although significant progress in speech synthesis has sometimes been based on intensive study of a single individual (e.g., Dennis Klatt), it is important for more generality to employ a variety of talkers in constructing test corpora. For example, we need to collect data from both males and females, several age groups, and several ethnic groups.

5. A 5th issue relates to the fact that articulation differs with speaking rate. For example, different segment types (e.g. vowels, consonants, and silent periods) increase and decrease differently as speaking rate changes. There are also differences in how precisely segments are articulated, and the nature and degree of coarticulation as speaking rate varies. Because our control strategies need to emulate these effects, our corpora must include a variety of speaking rates.

6. A 6th issue relates to the fact that different measurement techniques require different types of images. For example, some techniques can deal with ordinary facial images, while others require marking the lips with color for chromakey or affixing special markers for point tracking. We propose that a variety of talkers would be used as well as a number of visual recording techniques including usual full-face views and full face and mirror side views, and recordings with dot markers affixed to the face.

To achieve our validation goals we suggest that a number of new visual-auditory test corpora be created. These should include visual-auditory versions of some existing corpora, such as TIMIT. These corpora should be recorded on a high quality medium such as laser videodisk which includes aligned digital audio, orthography, phonetic, and paralinguistic labels.

We believe that these new corpora will be useful to the entire research community because they will allow multiple investigations of the same database for recognition by both humans and machines. Although some bimodal corpora may exist (e.g., a subset of the TIMIT sentences used by Goldschen [56]), distribution is currently limited for commercial reasons. We anticipate submission of these materials for distribution by the Linguistic Data Consortium. In Section 7.6.5 we will return to the issue of these materials as they relate to perceptual testing.

7.6.4 Comparing Measurements

Given the methods outlined in Section 7.6.2 for the analysis of facial information, we are faced with the problem of comparison between natural and synthetic information. To begin with, for any given measurement technique one might obtain measurements from the synthetic face in two ways. First, we could simply feed the synthetic images through the same analysis tools as those used on the natural face. Note, however, that 3D measurements based on multiple views would require

synthesis of these views. An alternate approach would be to simply make the measurements from the surfaces of the synthetic face.

Once we have obtained two sets of measurements, how should they be compared? It should be pointed out that often it may be of interest to simply examine the pattern of differences for single features at a time. For example, we might want to know whether the synthetic jaw rotates to the same degree as the natural one.

It is also of interest to evaluate the global agreement of natural and synthetic information. In a sense we can consider the features as locating test segments (e.g., phonemes or visemes) in a multidimensional space with our task being to compare the spatial arrangements of two sets of points. One possible way to do this would be to correlate the two sets of measures. Another possibility would be to use an error function, such as the absolute or squared difference in measures. For any of these methods, one would also want to examine both the instantaneous and dynamic agreement statistics.

A further complication in the comparison process could be that all measurements in a set might not cover the same range or be of equivalent importance. For example, lip width might be much more important than upper lip raising. Thus, we might want to weight the goodness metric differently for different measurement components. How would these weights be determined? One way would be to evaluate how important each measure is in providing optimal recognition. For example, Finn [47] used an algorithm to obtain the best weighting for recognition. A similar examination of various measures was used by Goldschen [56] in selecting which measurements would be used for recognition.

Other, higher level, analyses of the measured features might be of value in comparing facial articulations. For example, considering the measurements multidimensionally with multiple measurement sets, such as from different facial synthesizers, we might examine what weightings of the measurement dimensions bring the sets into closest agreement. As an example of another possible approach , Benoit, Lallouache, Mohamadi, and Abry [8] used dynamic clustering and correspondence analysis to classify the visemes used in a language. This could be applied to natural and synthetic measurements to assess their similarities.

Using these techniques we can arrive at some metrics of agreement between the behavior of natural and synthetic faces. However, while these metrics are of some value, they do not necessarily answer all of our validation concerns. First of all, different types of facial synthesis systems have been constructed for different purposes, and it may not be appropriate to judge them in the same way. For example, a system which simply uses analyzed features to drive synthesis may have less error than one which takes its input from text, but with much less flexibility. A second consideration is that the synthetic face may look nothing like a natural one. For example, we may be mapping human measurements to a dog's face [108]. This would of course result in large errors. Finally, how do we know that the measurements selected truly reflect the linguistic information conveyed facially? A partial answer was given just above - we can weight the measurements by how important they are for machine recognition. But this may be misleading because the features used by machines may differ from those used by human observers. Since the ultimate consumers of synthetic faces are humans, it is also essential to validate our work with human perceptual tests.

7.6.5 Perception Based Validation

In the perception-based approach to validation, we seek to evaluate how accurately human observers recognize the messages conveyed by the face and to what extent the perceptual results agree for

natural and synthetic faces. The latter analysis is quite important, since it is possible to create facial animations transmitting more information than natural speech (e.g., with subtitles).

An important assumption regarding our synthetic facial displays is that our representations are impoverished relative to natural facial behavior. Klatt [71] and Duffy and Pisoni [35] make a similar point for the case of synthetic auditory speech. They suggest that while in natural speech, linguistic information is conveyed by a number of redundant cues, synthetic speech lacks this redundancy because only a subset of the available cues are incorporated in the synthesis algorithms. It is probable that this lack of redundancy, as well as inaccuracies in segment target values, may affect our facial syntheses. We believe that the comparison of perception for natural and synthetic faces should help us pinpoint and remedy these problems.

7.6.6 Perceptual Paradigms

There are a number of issues regarding perceptual-based validation. The first issue, discussed earlier, concerns the size and type of the test stimuli. Different types of stimuli might engage different perceptual processes. For example, short nonsense words would presumably tap only low-level perceptual processes versus the involvement of lexical processes for real words. The presence of lexical information in the latter case could aid recognition by constraining the response alternatives. Similarly, when meaningful sentences are used, subjects can use syntactic and semantic constraints to improve performance. A couple of related issues in stimulus selection concern the complexity of the test units, and how the low level segments are specified. As an example of the first, in order to test segment coarticulation rules, it would be desirable to include words incorporating segment clusters (eg., consonant clusters and diphthongs) rather that the usual CVC words. As an example of the second problem, one might simply specify segment identities and durations analyzed from an actual human talker. However, since many systems will incorporate text-to-speech translation, it may be more appropriate to use the segment identities and durations derived from the translation module. To help pinpoint problems, we might actually want to test the systems while either including or bypassing various modules.

A second issue concerns the method of collecting responses from the observers. In some early tests (Diagnostic Rhyme Test (DRT) [139] and Modified Rhyme test [45, 61]), observers were presented with short words with a closed set of alternative responses (2 for DRT, 6 for MRT) which differed on the initial consonant. For example, if the test word was *bat*, the response alternatives might be *bat*, *cat*, *rat*, *sat*, *mat*, and *fat*. Although the closed form may be informative about the overall level of performance, it does not yield much information about the confusions made by the perceiver.

Although more difficult to score, a better approach is the open response method in which the observer simply reports the word heard. These response words can then be broken into constituent segments and compared with the segments actually presented, to form confusion matrices. For example, we can look at the responses in terms of actual initial consonant presented and perceived initial consonant. For sentence-length test material, this sort of analysis would have to be preceded by algorithms for alignment of the stimulus and response strings. Some examples of these algorithms are the NIST String Alignment and Scoring Program [63], and the sequence comparator of Bernstein, Demorest, and Eberhardt [11].

Given the confusion data for natural and synthetic faces, we can assess overall agreement and particular problem areas in our facial synthesis strategies. In addition to the direct comparison of confusion data, these strategies can be further analyzed using MDS techniques and the resulting

multidimensional spatial representations can be compared. By using techniques such as INDSCAL, multidimensional representations with common axes (though different dimensional weights) can be obtained. One may be able to then characterize different facial synthesis systems on the basis of these dimensional weights.

We should note that it is important to examine the confusions of different segment types and positions. For example, systems may do well on consonants but relatively poorly on vowels. Similarly, the transmission of consonants may vary depending on whether they occur in initial position, final position, or clusters.

In addition to analysis of accuracy and confusion data, a couple of other types of data may be of value. The first type is the response latency of the human observer. For example, for the auditory modality, even with the same level of accuracy, one may observe longer latencies for synthetic than natural speech, and different latencies for different synthesizers. One reason for this might be the relative lack of redundant information available in the synthetic forms [35]. For sentence-length materials we can also see differences in latencies for verification of sentence truth for natural and synthetic speech.

The second type of data is obtained by collecting quality ratings on natural and synthetic faces. For example, we can ask about whether the speech is too fast or slow, how easy it is to lipread, and how pleasing and realistic the face is. These ratings can then be compared.

Because synthetic visual speech may often be accompanied by auditory speech, it will be useful to test combined materials. To do that, we should use an efficient evaluation technique is to compare the visual speech intelligibility added to the auditory intelligibility, at various levels of acoustic degradation, utilizing a reference natural face and the model(s) we want to test. In that perspective, Le Goff, Guiard- Marigny, Cohen, and Benoit [54] compared the audiovisual intelligibility of the same corpus (18 phonetically constrained VCVCV sequences) with the same acoustic (naturally uttered) material. They used five different conditions of added noise, across four conditions of visual displays: no visual display (audio alone), the natural face of the speaker, the 3D model of the whole face (Parke modified by Cohen), and a 3D model of the lips. The results showed that the whole natural face restores the two-thirds of the missing auditory intelligibility when the acoustic transmission is degraded or missing; the facial model (tongue movements excluded) restores half of it; and the lip model restores a third.

It should be noted that the evaluation techniques presented here in the context of analysis of linguistic transmission can also be used for para-linguistic information (for example, emotion). In the auditory modality, for example, Cahn [22] analyzed confusions in the perception of the emotional content of sentences. In her study, five sentences were synthesized each presented in six emotions in a variety of random orders. The observers then identified which of the six were intended. Similar experiments can be carried out with synthetic faces at the sentence and smaller unit levels, and compared also with natural face transmission.

An issue not yet discussed concerns the selection of our perceptual observers. Who should they be? It may be that a number of different types are valuable. For example, in some early stages of development we may want to use experts such as trained FACS coders for evaluation of paralinguistic information transmission or expert lipreaders for evaluation of speech. These observers, in addition to analysis of their recognition levels and confusions, might be able to offer valuable qualitative insights about the synthesis systems. However, we would also want to test our systems with more naive observers who would be the typical end-users of our systems. The latter approach has been used in the majority of prior studies.

Another issue concerns internationalization. To allow the widest range of researchers to compare

their work, multilingual intelligibility tests are valuable. Some recent advances in this area are given in [10, 126, 7]

A final issue of concern when evaluating speech intelligibility with human observers is that of learning. This issue has received considerable attention in the context of training lipreading [52, 142], electrotactile [1], and unfamiliar speech distinctions [85]. The important point from these studies is that performance may change considerably with experience, so it is important to examine how well observers do with our synthetic faces both at first sight and later when they are more familiar with them.

8 Summary

[23]Faces communicate, entertain, express. We see a face framed by head, hair, hands, body. Participants at the workshop represented interests in all these, and evaluated the success of what they saw in different ways: as science, as communication, as entertainment. Such diversity simultaneously provided both vigor and challenge to the synthesis represented by this workshop.

8.1 Facial features

Even a passive face conveys a great deal of information. Social science tells us we see and react to species, gender, color, shape, wrinkles, freckles, hair, decorations (lipstick, mascara), and so on. Plastic surgeons carefully consider the structure of bone, skin, and muscles, while more generally circulation and innervation is also medically vital. As a face comes to life and begins to move, we observe the lips, teeth, and tongue for speech; the gestures of the eyes, head, and hands for dialog; and the flexing of muscles from forehead to neck for emotion. The available range of motions is remarkable. We puff cheeks, pout lips, stick out the tongue, raise an eyebrow, wink, yawn, scratch, and stare. Some actions are brief, some much slower, and some are variable, either consciously or unconsciously.

8.2 Motivation

Study of faces divided into two main categories, though not without overlap. Category one was analysis. This included speech recognition and lip reading both with and without sound, psychology, plastic surgery, and anesthesiology. The last item stems from evidence that EMG traces of facial muscles indicate levels of pain and consciousness even when a patient has been given anesthetics and muscle relaxants. Category two was synthesis. Here we had augmented speech and dialog synthesis, synthetic actors, cartoon animation, and virtual presence. There was much more freedom in synthetic faces, including cartoon images, cats, Martians, caricatures, and distortions, in addition to realistic natural faces. In fact, one synthetic face was reduced to lips alone, for speech studies.

Much of both analysis and synthesis was motivated by an interest in communication. Data was presented that implied noisy speech was significantly easier to understand when either a real or synthetic speaker was visible. Human-computer interfaces may be able to use that fact to advantage, by adding face synthesis to voice synthesis and recognition interfaces. It is not yet possible, however, to predict what response users will have to talking heads.

8.3 Final Remarks

Discussion groups examined three issues: data acquisition and recognition, modeling, and control. From these emerged a broad consensus on several points. First, it became apparent that many data bases already existed which were of wide interest. Efforts were instigated to make these available on the InterNet. Second, there was wide use of the Facial Action Coding System (FACS) for controlling emotional expression; this in spite of the fact that, more generally, there was such diversity in modeling and control that it would be challenging to present a unified framework. Third, tongue and timing were important to both analysis and synthesis, and much more data

---

[23]K. Shoemake

44

was needed on these features. Fourth, the level of detail available in FACS was not adequate to represent lips, tongue, and timing. It was clear that some augmented notation was needed, as many participants had already gone to the use of muscle level description or independent augmentation. Speech requires more detailed lip shapes than FACS provides, and even emotion requires more precise timing. Finally, there was consensus that standards could be helpful, both to facilitate data exchange and to support validation efforts. This last point is particularly important, as validation was seen as necessary and desirable in the work of almost everyone present.

In light of the number and diversity of interests represented at the workshop, it is perhaps remarkable that there was any consensus at all. Yet a common feeling expressed at the end was that this bringing together of colleagues, many previously unacquainted, had been enjoyable, educational, and valuable.

9 Goals Achieved and Recommendations

[24]Some of the achieved goals of the workshop are

- It produced a vocabulary list of signals that characterize faces and their motions.

- Due to the various and diverse applications there are, no unique standard seems appropriate but a minimum requirement of facial geometry and function is required.

- FACS is widely used in facial animation but needs more details regarding lip movements and the temporal data of muscle actions.

- Validation of facial models and controls is an important problem and different techniques were suggested.

- A videotape gathering works by the participants is being made.

In comparing the list of phenomena to be modeled to the list of modeling techniques, it is clear that researchers have been working effectively to address many of the fundamental modeling problems of facial animation. However, In spite of the array of advanced modeling technologies that have been introduced into the field, it is still a far from trivial task for an end user to create a model of a specific person's face and to make it speak or perform other complex behaviors. In order to enable practical applications of facial animation, five basic research tasks should be undertaken: understand application needs, develop a data description language, collect an extensive database, formalize and validate modeling techniques, and perform basic research into modeling techniques. As improved computing technology becomes available, new applications of facial modeling are becoming feasible and cost-effective. These research tasks are designed to encourage efficient development of these applications.

- **Understanding application needs**: None of the modeling topics is completely application independent since the application will at least determine to what level of detail certain phenomena are to be modeled. Researchers currently have few guidelines beyond common sense for determining which phenomena may safely be ignored in a simulation while still meeting the needs of the application. A detailed analysis of the potential applications for facial modeling would be a great benefit to researchers wishing to develop useful facial animation techniques.

---

[24]S. Platt and D. terzopoulos

45

- **Development of a language for describing facial data**: The facial modeling community currently lacks a standard method for recording facial models. The FACS system has been an invaluable tool for organizing research into facial actions, but it does not address geometric or physical properties of the face. In addition, FACS is not a computer standard, and implementations of FACS vary widely. It is likely that facial modeling will be able to benefit from the more general work in progress within the CAD and computer animation communities by providing face-specific extensions to standard file formats. Whatever approach to definition of a standard, it will no doubt undergo rapid evolution as new modeling techniques enter common use and therefore it should be built within a flexible framework. The format should be capable of representing information collected from living human faces in addition to purely synthetic faces in order to facilitate both validation and performance-driven hybrid models.

- **Collecting data**: Several of the validation techniques described above rely on the comparison of model generated data to data collected from living human faces. To facilitate validation and to provide insight into the structure and behavior of the face, the research community should be constructing a database of information about the faces of a wide range of human subjects. This task would include recording as much raw data as possible about the physical phenomena of the face. This effort would no doubt overlap with the effort described above to define a language for describing facial data, since it is through that language that the data would be stored and processed.

- **Formalization and validation of modeling techniques**: Now that facial modeling has been successfully applied to a range of application types, a standard and robust implementation of these techniques should be made available over the Internet for incorporation into new applications. This approach has proven extremely valuable for numerical algorithms and other computing tools. Together with the database of human subjects described above, this network resource should encourage application developers to incorporate sophisticated facial models in ways that might otherwise have been rejected as too difficult.

- **Further development of modeling techniques**: The field of facial modeling is still very new and promising as a field of research. The human face is perhaps one of the most difficult objects to model both because of its inherent complexity and because of the special attention human observers often pay to even the slightest details of shape, color, and motion. Many of the modeling techniques discussed in this section have only begun to be explored and others should be expanded and refined to incorporate advances from pure computer science and related engineering specialties.

FACS is the premier notation scheme being using for facial animation. However, FACS was designed as a recognitive scheme – defined actions were created as cognitively/visually distinct units rather than minimally generable units. In particular, (1) actions are imprecisely defined, and (2) actions combine in an unpredictable manner.

The imprecise definition is necessary with respect to the initial intent of FACS to notate general expression usage on human faces. However, describing precise animations can require the manipulation of the face at a sub-FACS level. The inability of FACS to exactly predict the result of an AU sequence is irrelevant to notators. A notator must be aware of what actions may not be currently visible; however, an animation control system must be able to explicitly state whether an action is visible. (Alternatively, facial region changes must be controllable at a sub-AU level, implying multiple levels of control even within the realm of FACS.) That FACS continues to function so well

despite these weaknesses, in a manner almost completely converse to its initial design, testifies to its flexibility and accuracy.

A general notation scheme for describing changes to the human face needs to be designed. The general philosophy of FACS (i.e., representable actions are described in parallel with a structural model of matching complexity) will be maintained. We will refer to this scheme as FACS+.

Control systems can be defined at several levels. The face suggests at least four: (1) geometric, (2) structural, (3) expressive, and (4) conversational. Geometric facial models manipulate the object at a purely physical level. Structural models represent the face in terms of active regions, taking a simplistic view of the underlying geometry. Expressive models work at a grosser feature level, animating faces based on its most obvious features. Conversational representation and control operate at an even higher level, dealing with emotional intent and general facial actions.

FACS operates at a structural level. Its ability to operate at a geometric level is limited due to its lack of definition at that low a level. Likewise, its ability to operate at higher levels has been well-researched, but exact mappings from intent to FACS AUs are specified in an ad hoc manner. FACS+ will also operate at a structural level. We assume other schemes will be used to operate on geometric and feature-based models; FACS+ will act as an intermediary between the two.

Issues involved in the replacement of FACS center on its relationship to lower and higher level schemes as well as extensions needed to more fully represent the face. Based on this, we have identified the following areas of concern when extending FACS to produce FACS+:

1. Downward links to physical model controls.

   - Mappings between FACS+ and lower-level controls need to be defined, as do methods of controlling the mappings.

2. Upward links to feature-based model controls.

   - Likewise, a control scheme operating at a gross feature level must be mappable and easily remappable into FACS+ controls.

3. Static definition of expression changes.

   - Actions need to be precisely (unambiguously) defined in an appropriate manner – a single primitive action should not possess variants or options. Note that feature-based model controls allow the creation of macro action options.

4. Expression dynamics.

   - Hooks to allow the expression of dynamics of action scripting should be present.

5. Muscle intensity.

   - Better control of action intensity needs to be added. Simply, a linear scale from 0 (no action) to 1 (full intensity) may be sufficient.

6. Extensibility.

   - The system should be extensible to allow for alternate physiological structures. This includes the redefinition of actions to account for deviant faces (scarring) as well as alternate architectures (non-human faces).

7. Fine/Coarse definition.

   - Certain areas of the face need finer definition. In particular, mouth and lip actions need improvement. Other FACS actions based on timing (blink, wink, etc.) or intensity (various lip pull actions) need to be redefined to eliminate temporal and intensity-based components.

8. Tongue action.

   - Actions of the tongue need to be developed. These are limited to gross actions (position and shape of the tongue), and do not include interactions between the tongue and other facial structures.

9. Interactions.

   - A method of further defining interactions between structures needs to be developed (e.g., cheek thrust). Although these are handled at a geometric level by simple interaction tests and constraint propagation, a FACS+ representation will provide a necessary link between the physically based models and high-level feature models.

10. External interactions.

   - Likewise, a general method needs to be defined to handle the effects on facial structures due to external influences. These include but are not limited to gravity on slack faces and cheek puffing.

Attempts are already underway aimed at extending FACS by defining methods to automate FACS-coding, and to extend and improve the modeling, especially within the context of simulations, animations and human-machine interaction. Ekman and Sejnowski [39] are at present developing a neural net approach for recognizing FACS AUs. Yacoob and Davis [152] have developed a facial expression recognition system based on FACS. Essa and Pentland [42, 44] are concentrating on extending the FACS model to FACS+ by observing real people making expressions and extracting spatial and temporal information from video to describe facial motion.

10   List of Participants

Ken Anjyo
Hitachi Research Laboratory
Hitachi, Ltd.
comE-mail: anjyo@hrl.hitachi.co.jp
Phone: +81-3-5485-1451   FAX: +81-3-5485-1457

Norman Badler
University of Pennsylvania
Department of Computer and Information Science
E-mail: badler@central.cis.upenn.edu
Phone: (215) 898-5862 FAX: (215) 898-0587

Tripp Becket

48

University of Pennsylvania
Department of Computer and Information Science
E-mail: becket@graphics.cis.upenn.edu
Phone: (215) 898-3587 FAX: (215) 898-0587

Henry Bennett
Department of Anesthesiology
UC Davis Medical Center
2315 Stockton Blvd.,
Sacramento, CA 95817

Christian Benoit
CNRS-INPG-Universite Stendhal
ICP - Institute of Speech Communication, Grenoble, France
E-mail: benoit@icp.grenet.Fr
Phone: (+33) 76 82 43 36 FAX: (+33) 76 82 43 35

Justine Cassell
University of Pennsylvania
Department of Computer and Information Science
E-mail: justine@central.cis.upenn.edu
Phone: (215) 573-2821 FAX: (215) 898-0587

Michael M. Cohen
University of California - Santa Cruz
Program in Experimental Psychology
E-mail: mmcohen@dewi.ucsc.edu
Phone: (408) 459-2655 FAX: (408) 459-3519

Irfan Essa
MIT Media Lab
Perceptual Computing Section
E-mail: irfan@media.mit.edu
Phone: (617) 253-3891 FAX: (617) 253-8874

Oscar Garcia
George Washington University
E-mail: garcia@seas.gwu.edu
Phone: (703) 306-1928

Alan Goldschen
George Washington University
E-mail: ajg@seas.gwu.edu
Phone: (301) 596-1474

Joseph Hager

University of California
E-mail: joehager@ucsfvm.ucsf.edu
Voice: (415) 476-7207 Voicemail: (415) 824-3023

John Hestenes
Biomedical Engineering and Science Institute
Drexel University
E-mail: jhestene@ece.drexel.edu

Prem Kalra
University of Geneva
MIRAlab
E-mail: KALRA@uni2a.unige.ch
Phone: (+41) 22 705 7766 FAX: (+41) 22 320 2927

Alan Kaplan
AT&T Bell Laboratories
Murray Hill, NJ
E-mail: aek@research.att.com
Phone: (908) 582-7542 FAX: (908) 582-5857

Tsuneya Kurihara
Central Research Laboratory
Hitachi, Ltd.
E-mail: kurihara@crl.hitachi.co.jp
Phone: +81-423-23-1111 FAX: +81-423-27-7699

Pete Litwinowicz
Apple Computers
E-mail: litwinow@apple.
Phone: (408) 974-1752 FAX: (408) 862-5520

Pamela Mason
Patient Comfort Inc.
127 Marguerite Lane
Cloverdale, CA 95425
Phone: (707) 894-3013

Marilyn Panayi
Visiting Scholar - University of Delaware
Department of Education
E-mail: panayi@asel.udel.edu

Fred Parke
IBM, /LOB Technology, 9462
11400 Burnet Road

Austin, TX 78758
E-mail: parke@futserv.austin.ibm.com

Manjula Patel
University of Bath
School of Mathematical Sciences
E-mail: mp@maths.bath.ac.uk
Phone: (+44) 225 826183 FAX: (+44) 225 826492

Catherine Pelachaud
University of Pennsylvania
Department of Computer and Information Science
E-mail: pelachau@graphics.cis.upenn.edu
Phone: (215) 898-1976 FAX: (215) 898-0587

Eric Petajan
AT&T Bell Laboratories
E-mail: edp@allegra.att.com
Phone: (908) 582-3160

Steven Pieper
University Dartmouth
Medical Media Systems
E-mail: Stevie.Pieper@dartmouth.edu
Phone: (603) 646-2623 FAX: (603) 646-3805

Stephan Platt
University of Pennsylvania
Department of Computer and Information Science
E-mail: platt@graphics.cis.upenn.edu

Scott Prevost
University of Pennsylvania
Department of Computer and Information Science
E-mail: prevost@linc.cis.upenn.edu
Phone: (215) 898-9511 FAX: (215) 898-0587

David M. Roy
Visiting Scholar - University of Delaware
Applied Science and Engineering Laboratories
E-mail: roy@asel.udel.edu
Phone: (302) 651-6830 FAX: (302) 651-6895

Agnes Saulnier
INA - FRANCE
E-mail: saulnier@ina.fr

Phone: (+33) 1 45 80 2041 FAX: (+33) 1 49 83 2582

Mark Steedman
University of Pennsylvania
Department of Computer and Information Science
E-mail: steedman@linc.cis.upenn.edu
Phone: (215) 898-2012 FAX: (215) 898-0587

Demetri Terzopoulos
University of Toronto
E-mail: dt@vis.toronto.edu
Phone: (416) 978-7777 FAX: (416) 978-1455

Marie-Luce Viaud
University of Pennsylvania
Department of Computer and Information Science
E-mail: luce@graphics.cis.upenn.edu
E-mail: luce@ina.fr
Phone: (215) 898-1976 FAX: (215) 898-0587

Carol Wang
University of Calgary
Department of Computer Science
E-mail: cpsc.ucalgary.ca

Keith Waters
Digital Equipment Corporation
Cambridge Research Laboratory
E-mail: waters@crl.dec.com

Hussein Yahia
INRIA Rocquencourt
E-mail: hussein@bora.inria.fr
Phone: (33-1) 39 63 53 57 FAX: (33-1) 39 63 53 30

References

[1] J. Allen, M.S. Hunnicutt, and D. Klatt. *From text to speech: The MITalk system*. Cambridge University Press, Cambridge, MA, 1987.

[2] K. Anjyo, Y. Usami, and T. Kurihara. A simple method for extracting the natural beauty of hair. *Computer Graphics*, 26(2):111–120, July 1992.

[3] Ascension Technology Corp., P.O. Box 527, Burlington, Vermont 05402, Tel: (802)-860-6440.

[4] A. Azarbayejani, T. Starner, B. Horowitz, and A.P. Pentland. Visually controlled graphics. *IEEE Trans. Pattern Analysis and Machine Vision*, 15(6), June 1993.

[5] H. Bennett. F.A.C.E.: A sensitive and specific monitor for adequacy of anesthesia. In P. Segel, G. Winograd, , and B. Bonke, editors, *Memory and Awareness in Anesthesia*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] C. Benoit, A. Adjoudani, O. Angola, T. Guiard-Marigny, and B. Le Goff. Perception, analysis and synthesis of talking lips. In *Proc. of the IMAGINA '94 Conference*, pages 142–163. Institut National de l'Audiovisuel, France, Feb. 1994.

[7] C. Benoit, M. Grice, and V. Hazan. The sus test: A method for the assessment of text-to-speech synthesis intelligibility. *Speech Communication*, under revision.

[8] C. Benoit, T. Lallouache, T. Mohamadi, and C. Abry. A set of french visemes for visual speech synthesis. In *Talking machines: Theories, Models, and Designs*, pages 485–504. Elsevier North-Holland, Amsterdam, 1992.

[9] C. Benoit and L.C.W. Pols. On the assessment of synthetic speech. In *Talking machines: Theories, Models, and Designs*, pages 435–442. North Holland, Amsterdam, 1992.

[10] C. Benoit, A. van Erp, M. Grice, V. Hazan, and U. Jekosh. Multilingual synthesizer assessment using semantically unpredictable sentences. In *Proceedings of the 2nd EUROSPEECH Conference*, pages 633–636, Paris, France, 1989.

[11] L.E. Bernstein, M.E. Demorest, and S.P. Eberhardt. A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment. *Journal of the Acoustical Society of America*, submitted.

[12] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *International Joint Conference of Speech and Signal Processing*, pages 557–560, Minneapolis, MN, 1993.

[13] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proceedings of the Int. Conf. on Acoustics Speech and Signal Processing (IEEE-ICASSP)*, Adelaide, Australia, 1994.

[14] C. Bregler and S. Omohundro. Surface learning with applications to lip-reading. In *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufman, Palo Alto, CA, 1994.

[15] N. M. Brooke and E. D. Petajan. Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. In *Proceedings of the International Conference on Speech Input/Output: Techniques and Applications*, pages 104–109, London, UK, 1986. Institution of Electrical Engineers.

[16] N.M. Brooke. *The Structure of Multimodal Dialogue*, chapter Visible speech signals: Investigating their analysis, synthesis, and perception. Holland: Elsevier Science Publishers, 1989.

[17] P. Brun, I. Parienti, and P. Serres. *Les cahiers de médecine esthétique*. Solal Editions, Paris, 1987.

[18] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, October 1993.

[19] J. Buhmann, J. Lange, and C. von der Malsburg. Distortion invariant object recognition by matching hierarchically labeled graphs. In *IJCNN International Conference on Neural Networks, vol. 1*, pages 155–159, Washington, DC, 1989.

[20] P.E. Bull and R. Brown. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 16, 1977.

[21] Abry C. and Boe L.-J. Laws for lips. *Speech Communication*, 5:97–104, 1986.

[22] J. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technolog, 1989.

[23] J. Cassell and D. McNeil. Non-verbal imagery and the poetics of prose. *Poetics Today*, 12(3):375–404, 1990.

[24] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graghics Annual Conferences Series*, pages 413–420, 1994.

[25] M.A. Cathiard, G. Tiberghien, A. Cirot-Tseva, M.T. Lallouache, and P. Escudier. Visual perception of anticipatory rounding during acoustic pauses: A cross-linguistic study. In *Proceedings of the 12th Int. Congress of Phonetic Sciences*, pages 50–53, Aix-en-Provence, France, 1991.

[26] C.S. Choi, H. Harashima, and T. Takebe. Highly accurate estimation of head motion and facial action information on knowledge-based image coding. Technical Report PRU90-68, I.E.I.C.E.J., 1990.

[27] M. M. Cohen and D. W. Massaro. Synthesis of visible speech. *Behavioral Research Methods and Instrumentation*, 22(2):260–263, 1990.

[28] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, Tokyo, 1993. Springer-Verlag.

[29] W.S. Condon and W.D. Osgton. Speech and body motion synchrony of the speaker-hearer. In D.H. Horton and J.J. Jenkins, editors, *The Perception of Language*, pages 150–184. Academic Press, 1971.

[30] I. Craw, H. Ellis, and J.R. Lishman. Automatic extraction of face-features. *Pattern Recognition Letters*, 5:183–187, 1987.

[31] Cyberware Laboratory Inc., Monterey. *4020/(RGB 3D) Scanner with color digitizer*, 1990.

[32] H. Delinguette, G. Subsol, S. Cotin, and J. Pignon. A craniofacial surgery simulation testbed. Technical Report 2199, INRIA, France, February 1994.

[33] S. DiPaola. Implementation and use of a 3d parameterized facial modeling and animation system. In *Vol 22: State of the Art in Facial Animation*, pages 20–33. ACM Siggraph'89 Course Notes, 1989.

[34] S. Dubin, J. Nissanov, S. Zietz, B. Schrope, R. Morano, and R. Hananiah. Bioengineering approach to non-invasive measurement of body composition. In *Rocky Mountain Bioengineering Symposium*, pages 21–23, April 1994.

[35] S.A. Duffy and D.B. Pisoni. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. Technical Report 17, Research on Speech Perception, Indiana University, Department of Psychology, 1991.

[36] P. Ekman and W. Friesen. *Unmasking the Face: A guide to recognizing emotions from facial clues*. Prentice-Hall, NY, 1975.

[37] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Inc., Palo Alto, CA, 1978.

[38] P. Ekman, T.S. Huang, T.J. Sejnowski, and J.C. Hager. Final report to NSF of the planning workshop on facial expression understanding. Technical report, NSF, July 30 to August 1 1992.

[39] P. Ekman and T. Sejnowski. personal communication, 1994.

[40] M. Elson. "Displacement" facial animation techniques. In *Vol 26: State of the Art in Facial Animation*, pages 21–42. ACM Siggraph'90 Course Notes, Dallas Convention Center, August 6th–10th 1990.

[41] A. Emmett. Digital portfolio: Tony de Peltrie. *Computer Graphics World*, 8(10):72–77, October 1985.

[42] I. A. Essa. *Analysis, Interpretation, and Synthesis of Facial Expressions*. PhD thesis, MIT, Media Laboratory, Cambridge, MA, 1994.

[43] I. A. Essa, T. Darrell, and A. Pentland. Tracking facial motion. In *Proceedings of IEEE Workshop on Nonrigid and Articulated Motion*, 1994.

[44] I.A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. *Proceedings of Computer Vision and Pattern Recognition (CVPR 94)*, pages 76–83, 1994.

[45] G. Fairbanks. Test of phonemic differentiation: The rhyme test. *Journal of the Acoustical Society of America*, 30:596–600, 1958.

[46] E.K. Finn and A.A. Montgomery. Automatic optically based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.

[47] K. E. Finn. *An Investigation of Visible Lip Information to be Used in Automated Speech Recognition*. PhD thesis, Georgetown University, Washington, DC, 1986.

[48] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1), January 1973.

[49] D.R. Forsey and R.H. Bartels. Hierarchical bspline refinement. *Computer Graphics*, 22(4):205–212, May 1990.

[50] V. Fromkin. Lip positions in American-English vowels. *Language and Speech*, 7:215–225, 1964.

[51] O.N. Garcia, A.J. Goldschen, and E.D. Petajan. Feature extraction for optical automatic speech recognition or automatic lipreading. Technical Report GWU-IIST-92-32, Institute for Information Science and Technology, George Washington University, 1992.

[52] A.T. Gesi, D.W. Massaro, and M.M. Cohen. Discovery and expository methods in teaching visual consonant- and word-identification. *Journal of Speech and Hearing Research*, 35:1180–1188, 1992.

[53] S. Glenn. VActor animation creation system. In *ACM SIGGRAPH '93 Visual Proceedings*, page 223, 1993. SimGraphics Engineering Corporation.

[54] B. Le Goff, T. Guiard-Marigny, M.M. Cohen, and C. Benoit. Real-time analysis-synthesis and intelligibility of talking faces. In *Proc. of the 2nd ESCA/IEEE workshop on Speech Synthesis*, pages 53–56, New Paltz, NY, 1994.

[55] P. Goldenthal. Posing and judging facial expressions of emotion: The effects of social skills. *Journal of Social and Clinical Psychology*, 3(3):325–338, 1985.

[56] A. J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading*. PhD thesis, George Washington University, 1993.

[57] P. Griffin and N. Hoot. The FERSA project for lip-sync animation. Technical report, Department of Interactive systems CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands, 1993.

[58] T. Guiard-Marigny, A. Adjoudani, and C. Benoit. A 3-d model of the lips for visual speech synthesis. In *Proc. of the 2nd ESCA/IEEE workshop on Speech Synthesis*, pages 49–52, New Paltz, NY, 1994.

[59] P. Hanrahan and W. Krueger. Reflection from layered surfaces due to subsurface scattering. *Computer Graphics Annual Conference Series*, pages 165–174, 1993.

[60] D.R. Hill, A. Pearce, and B. Wyvill. Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer*, 3:277–289, 1988.

[61] A. S. House, C. E. Williams, M. H. Hecker, and K. D. Kryter. Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37:158–166, 1965.

[62] P. L. Jackson, A. A. Montgomery, and C. A. Binnie. Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, 19(4):796–812, Dec 1976.

[63] S. Janet. *NIST String Alignment and Scoring Program*. National Institute of Standards and Technology, 1988.

[64] S. Kaiser and T. Wehrle. Automated coding of facial behavior in human-computer interactions with FACS. *Journal of Nonverbal Behavior*, 16:2, 1992.

[65] P. Kalra, E. Gobbetti, N. Magnenat-Thalmann, and D. Thalmann. A multimedia testbed for facial animation control. In T.S. Chua and T.L. Kunii, editors, *International Conference of Multi-Media Modeling, MMM'93*, pages 59–72, Singapore, Nov 9-12, 1993.

[66] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. Smile: A multilayered facial animation system. In T.L. Kunii, editor, *Modeling in Computer Graphics*. Springer-Verlag, 1991.

[67] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proceedings of the First International Conference on Computer Vision*. IEEE Computer Society Press, 1987.

[68] M. Kass, A. Witkin, and D. Terzopoulous. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[69] A. Kendon. Movement coordination in social interaction: Some examples described. In Weitz, editor, *Nonverbal Communication*. Oxford University Press, 1974.

[70] R.D. Kent and F.D. Minifie. Coarticulation in recent speech production models. *Journal of Phonetics*, 5:115–135, 1977.

[71] D.H. Klatt. A review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82:737–793, 1987.

[72] J. Kleiser. A fast, efficient, accurate way to represent the human face. In *Vol 22: State of the Art in Facial Animation*, pages 20–33. ACM Siggraph'89 Course Notes, 1989.

[73] Kleiser-Walczak. Sextone for President. *ACM SIGGRAPH Video Review*, vol. 38/39, 1988. Kleiser Walczak Construction Comp.

[74] H. Knudsen and L. Muzekari. *The Effect of Verbal Statements of Context on Facial Expressions of Emotion*. 1980.

[75] T. Kurihara and K. Arai. A transformation method for modeling and animation of the human face from photographs. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 45–58. Springer-Verlag, 1991.

[76] M.T. Lallouache. Acquisition et traitement de contours labiaux. In *18ièmes, JEP du GCP du GALF, Montreal*, pages 27–28, 1990.

[77] M.T. Lallouache and C. Worley. Saisie, édition et traitement d'images et de signaux articulatoires des lèvres et de la machoire. *Journal d'Acoustique*, 1:215–220, 1990.

[78] W. Larrabee. A finite element model of skin deformation. I. Biomechanics of skin and soft tissue: A review. *Laryngoscope*, 96:399–405, 1986.

[79] A. LeBlanc, P. Karla, N. Magnenat-Thalmann, and D. Thalmann. Sculpting with the "ball and mouse" metaphor. In *Proc. Graphics Interface '91*, Calgary, Canada, 1991.

[80] Y. Lee, D. Terzopoulos, and K. Waters. Constructing physics-based facial models of individuals. In *Graphics Interface '93*, pages 1–8, Toronto, ON, May 1993.

[81] J.P. Lewis and F.I. Parke. Automated lipsynch and speech synthesis for character animation. In *Proceedings Human Factors in Computing Systems and Graphics Interface '87*, pages 143–147, 1987.

[82] H. Li, P. Roivainen, and R. Forchheimer. 3-d motion estimation in model-based facial image coding. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):545–555, June 1993.

[83] B. Lindblom and J. Sundberg. Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50:1166–1179, 1971.

[84] A. Lofqvist. Speech as audible gestures. *Speech Production and Speech Modeling*, pages 289–322, 1990.

[85] J.S. Logan, S.E. Lively, and D.B. Pisoni. Training Japanese listeners to identify /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89:874–886, 1989.

[86] Logitech Inc., 6505 Kaiser Drive, Fremont, CA 94555, Tel: (510)-795-8500.

[87] N. Magnenat-Thalmann, N.E. Primeau, and D. Thalmann. Abstract muscle actions procedures for human face animation. *Visual Computer*, 3(5):290–297, 1988.

[88] N. Magnenat-Thalmann and D. Thalmann. The direction of synthetic actors in the film: Rendez-vous à Montrèal. *IEEE Computer Graphics and Applications*, pages 9–19, December 1987.

[89] N. Magnenat-Thalmann and D. Thalmann. Synthetic actors. In *Vol 22: State of the Art in Facial Animation*, pages 145–152. ACM Siggraph'89 Course Notes, 1989.

[90] K. Mase and A. Pentland. Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22, 1991.

[91] K. Matsuoka, T. Furuya, and K. Kurosu. Speech recognition by image processing of lip movements. *Trans. on Soc. Instru. and Cont. Eng*, 22(2):191–198, 1986.

[92] K. Meyer, H.L. Applewhite, and F.A. Biocca. A survey of position trackers. *Presence*, 1(2):173–200, Spring 1992.

[93] A. A. Montgomery and P. L. Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America*, 73:2134–2144, 1983.

58

[94] S. Morishima and H. Harashima. A media conversion from speech to facial image for intelligent man-machine interface. *IEEE Journal on Selected Areas in Communications*, 9(4), 1991.

[95] S. Morishima and H. Harashima. Speech-to-image media conversion based on VQ and neural network. In *Proceedings of ICASSP91, M10.11*, pages 2865–2868, 1991.

[96] S. Muraki. Volumetric shape description of range data using "blobby model". *Computer Graphics*, 25(4):227–236, July 1991.

[97] L. Muzekari and H. Knudsen. Effect of context on perception of emotion among psychiatric patients. *Perceptual and Motor Skills*, 62(1):79–84, 1986.

[98] M. Nahas, H. Huitric, and M. Saintourens. Animation of a B-spline figure. *The Visual Computer*, 3(5):272–276, March 1988.

[99] K. Ohmura, A. Tomono, and Y. Kobayashi. Method of detecting face direction using image processing for human interface. *SPIE, Visual Communications and Image Processing '88*, 1001:625–632, 1988.

[100] A. Paouri, N. Magnenat-Thalmann, and D. Thalmann. Creating realistic three-dimensional human shape characters for computer-generated films. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 45–58. Springer-Verlag, 1991.

[101] F. I. Parke. Parameterized models for facial animation - revisited. In *Vol 22: State of the Art in Facial Animation*, pages 43–56. ACM Siggraph'89 Course Notes, 1989.

[102] F.I. Parke. Computer generated animation of faces. Master's thesis, University of Utah, Salt Lake City, UT, June 1972. UTEC-CSc-72-120.

[103] F.I. Parke. *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Salt Lake City, UT, 1974.

[104] F.I. Parke. A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, 1(1):1–4, 1975.

[105] F.I. Parke. A parameterized model for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–70, 1982.

[106] M. Patel. *Making FACES*. PhD thesis, School of Mathematical Sciences, University of Bath, Bath, Avon, UK, 1991.

[107] M. Patel and P.J. Willis. FACES—The facial animation, construction and editing system. In *Eurographics '91*, pages 33–45, Austria, 1991.

[108] E.C. Patterson, P.C. Litwinowicz, and N. Greene. Facial animation by spatial mapping. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, Tokyo, 1991. Springer-Verlag.

[109] A. Pearce, B. Wyvill, G. Wyvill, and D. Hill. Speech and expression: A computer solution to face animation. In *Graphics Interface '86*, 1986.

[110] C. Pelachaud. *Communication and Coarticulation in Facial Animation*. PhD thesis, Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA, 1991.

[111] C. Pelachaud, N.I. Badler, and M. Steedman. Linguistic issues in facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 15–30. Springer-Verlag, 1991.

[112] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech from phonemic, intonational, and affectual representations. *Cognitive Science*, in press.

[113] C. Pelachaud, M.L. Viaud, and H. Yahia. Rule-structured facial animation system. In *IJCAI '93*, 1993.

[114] A. Pentland and K. Mase. Lipreading: Automatic visual recognition of spoken words. In *Proc. Image Understanding and Machine Vision*. Optical Society of America, June 12-14, 1989.

[115] A. Pentland, B. Moghaddam, and T. Starner. *Computer Vision and Pattern Recognition Conference*, chapter View-based and modular eigenspaces for face recognition, pages 84–91. IEEE Computer Society, 1994.

[116] E. D. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition. In *CHI 88*, pages 19–25, 1988.

[117] E.D. Petajan. Automatic lipreading to enhance speech recognition. In *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, November 1984.

[118] E.D. Petajan. Automatic lipreading to enhance speech recognition. In *IEEE Computer society conference on computer vision and pattern recognition*, pages 40–47, June 1985.

[119] S. Pieper and D. Zeltzer. A biologically inspired model of human facial tissue for computer animation. In *Vol 22: State of the Art in Facial Animation*, pages 71–124. ACM Siggraph'89 Course Notes, 1989.

[120] S.D. Pieper. More than skin deep: Physical modeling of facial tissue. Master's thesis, Massachusetts Institute of Technology, Media Arts and Sciences, 1989.

[121] S.D. Pieper. *CAPS: Computer-Aided Plastic Surgery*. PhD thesis, Massachusetts Institute of Technology, Media Arts and Sciences, September 1991.

[122] S.M. Platt. A system for computer simulation of the human face. Master's thesis, University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, PA, 1980.

[123] S.M. Platt. *A Structural Model of the Human Face*. PhD thesis, University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, PA, 1985.

[124] S.M. Platt and N.I. Badler. Animating facial expressions. *Computer Graphics*, 15(3):245–252, 1981.

[125] Polhemus Navigation Sciences, 1 Hercules Drive, Colchester, VT 05446, Tel: (802)-655-3159.

[126] L.C.W. Pols and SAM Partners. Multilingual synthesis evaluation methods. In *Proceedings of the 2nd International Conference on Spoken Language Processing*, pages 181–184, Banff, Alberta, Canada, 1989.

[127] D. Reisfeld and Y. Yeshurun. Robust detection of facial features by generalized symmetry. *IEEE*, 1992.

[128] A.E. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27, 1964.

[129] T.W. Sederberg and S.R. Parry. Free form deformation of solid geometric models. *Computer Graphics*, 20(4):151–160, August 1986.

[130] SimGraphics Engineering Corp., 1137 Huntington Drive, South Pasadena, California 91030, Tel: (213)-255-0900.

[131] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. *IJCNN*, June 1992.

[132] S. Tamura, N. Kajimi, K. Okazaki, H. Mitsumoto, H. Kawai, and Y. Fukui. Lip contour extraction-complement and tracing by using energy function and optical flow. *Papers of Technical Group on Pattern Recognition and Understanding. I.E.I.C.E.*, PRU89-20:9–16, 1989.

[133] D. Terzopoulos and K. Waters. Physically-based facial modeling, analysis, and animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[134] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 569–579, 1993.

[135] D. Terzopoulous and K. Waters. Techniques for realistic facial modeling and animation. In M. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, Tokyo, 1991. Springer-Verlag.

[136] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[137] M.W. Vannier, T. Pilgram, G. Bhatia, and B. Brunsden. Facial surface scanner. *IEEE Computer Graphics and Applications*, 11(6):72–80, 1991.

[138] M.L. Viaud and H. Yahia. Facial animation with wrinkles. In *3nd Workshop on Animation, Eurographics '92*, Cambridge, 1992.

[139] W.D. Voiers. Performance evaluation of speech processing devices: Diagnostic evaluation of speech intelligibility. Technical Report Contract AF19(628)-4987, AFCLR-67-0101, AF Cambridge Research, 1967.

[140] Votrax, Div. of Federal Screw Works. *User's Manual*, 1981.

[141] C.T. Waite. The Facial Action Control Editor, FACE: A parametric facial expression editor for computer generated animation. Master's thesis, Massachusetts Institute of Technology, Media Arts and Sciences, Cambridge, February 1989.

[142] B. Walden, R. Prosek, A. Montgomery, C. K. Scherr, and C. J. Jones. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20:120–145, 1977.

[143] C.L.Y. Wang. Langwidere: A hierarchical spline based facial animation system with simulated muscles. Master's thesis, University of Calgary, Department of Computer Science, Calgary, AB, October 1993.

[144] K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics*, 21(4):17–24, July 1987.

[145] K. Waters. *The Computer Synthesis of Expressive Three-Dimensional Facial Character Animation*. PhD thesis, Middlesex University, Faculty of Art and Design, Cat Hill Barnet Herts, EN4 8HT, June 1988.

[146] K. Waters. A dynamic model of facial tissue. In *Vol 22: State of the Art in Facial Animation*, pages 145–152. ACM Siggraph'89 Course Notes, 1989.

[147] K. Waters and D. Terzopoulos. A physical model of facial tissue and muscle articulation. *Proceedings of the First Conference on Visualization in Biomedical Computing*, pages 77–82, May 1990.

[148] K. Waters and D. Terzopoulos. Modeling and animating faces using scanned data. *Journal of Visualization and Animation*, 2(4):123–128, December 1991.

[149] K. Waters and D. Terzopoulos. The computer synthesis of expressive faces. *Phil. Trans. R. Soc. Lond. B*, 355(1273):87–93, Jan 1992.

[150] P.M. Will and K.S. Pennington. Grid coding: A preprocessing technique for robot and machine vision. *Artificial Intelligence*, 2:319–329, 1971.

[151] L. Williams. Performance-driven facial animation. *Computer Graphics*, 24(4):235–242, August 1990.

[152] Y. Yacoob and L. Davis. *Computer Vision and Pattern Recognition Conference*, chapter Computing spatio-temporal representations of human faces, pages 70–75. IEEE Computer Society, 1994.

[153] A. Yuille, D. Cohen, and P. Hallinan. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.

[154] A.L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.

[155] A.L. Yuille, D.S. Cohen, and P.W. Hallinan. Feature extraction from faces using deformable templates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'89)*, pages 104–109, San Diego, California, June 1989. IEEE Computer Society Press.

[156] J.P. Zerling. *Aspects articulatoires de la labialité vocalique en francais*. PhD thesis, Institute of Phonetics, Strasbourg, France, 1990.

[157] I. Zlokarnik. Experiments with an articulatory speech recognizer. In *Eurospeech '93*, Berlin, Sept. 1993.