Final Report To NSF of the Planning Workshop on Facial Expression Understanding

July 30 to August 1, 1992

Edited by
Paul Ekman
Thomas S. Huang
Terrence J. Sejnowski
Joseph C. Hager

Note: Edited for Web browser presentation 7/97 & 4/00 by J.H.; printed edition 3/93 can be requested from Human Interaction Lab, LPPI Box 0984, University of California, San Francisco, CA 94143.

This workshop was supported by the Interactive Systems Program, the Robotics and Machine Intelligence Program, the Social Psychology Program, and the Division of Behavioral and Cognitive Sciences, all entities of the National Science Foundation. The views expressed in this report represent the views of the participants, not of the National Science Foundation.

Contents

I. EXECUTIVE SUMMARY by P. Ekman and T. J. Sejnowski
I-A. Introduction
I-B. Background About the Importance of Facial Expression
I-C. Progress and Obstacles in Measuring Facial Expression
I-D. The Opportunity
I-E. Recommendations
I-F. Organization of the Report
II. OVERVIEW OF THE WORKSHOP by P. Ekman
II-A. Organization of the Workshop
II-B. Scope of the Workshop
II-C. Goals achieved
II-D. Workshop Schedule
II-E. Participants
Invited Participants:
Other Workshop Participants:
Post-Workshop Contributor:
III. TUTORIAL SUMMARIES
III-A. The Psychology and Neuroanatomy of Facial Expression by J. Cacioppo, J. C. Hager, and P. Ekman
Facial Signal Systems
Techniques for Measuring the Rapid Facial Signals
Evidence About Which Facial Actions Signal Which Emotions
The Neuroanatomy of Facial Movement
Facial Data Base
III-B. Computer Vision and Face Processing by T. S. Huang, P. Burt, and K. Mase
Introduction
Representation

Reconstruction
Object Recognition
Motion Analysis and Face Processing
Optic Flow Approach to Face Processing
III-C. Neural Networks and Eigenfaces for Finding and Analyzing Faces by A. Pentland and T. Sejnowski
Introduction
Important Subproblems
Major Approaches
Biological Foundation of Face Processing
Face Detection
Face Recognition
Tracking Faces
Lip reading
Neural Networks
Backpropagation of Errors
Sex Recognition
Expression Recognition
III-D. Special Hardware For Face Processing by D. Psaltis
Analog VLSI
Optics
IV. REPORTS FROM THE PLANNING BREAKOUT GROUPS
IV-A. Breakout Group on Basic Science
Basic Science for Understanding Facial Expression by R. Davidson, J. Allman, J. Cacioppo, P. Ekman, W. Friesen, J. C. Hager, and M. Phillips
Recommendations on Basic Research Needs
Infrastructure Recommendations
Next Steps towards Automating Facial Measurement
Automating Other Relevant Data Sources:
IV-B. Breakout Group on Sensing and Processing
Sensing and Processing by A. Yuille, A. Pentland, P. Burt, G. Cottrell, O. Garcia, H. Lee, K. Mase, and T. Vetter
Introduction and Overview
Sensing and Environments.
Detection of Faces
Feature Extraction from Static Images
Feature Extraction from Image Sequences
Lip Reading
Expression Recognition
IV-C. Breakout group on Modeling and Databases
Computer-Based Facial Expression Models and Image Databases by F. Parke, D. Terzopoulos, T. Sejnowski, P. Stucki, L. Williams, D. Ballard, L. Sadler, and J. C. Hager
Introduction
State of the Art in 3D Modeling
Facial Databases
Research Directions and Goals
V. RECOMMENDATIONS by J. C. Hager
Basic Research on the Face that Answers These Crucial Questions:
Infrastructure Resources that Include the Following:

Tools for Processing and Analyzing Faces and Related Data:
Training and Education for Experienced and Beginning Investigators:
VI. BENEFITS by B. Golomb and T. J. Sejnowski
Commercial Applications
Computer Sciences
Basic Science Research
Medicine
Unforeseen Benefits
VII. REFERENCES.

NSF Report - Facial Expression Understanding

I. EXECUTIVE SUMMARY

Paul Ekman and Terrence J. Sejnowski

I-A. Introduction

Grand challenges like space exploration and weather prediction are expanding human frontiers, but the grandest challenge is the exploration of how we as human beings react to the world and interact with each other. Faces are accessible "windows" into the mechanisms which govern our emotional and social lives. The technological means are now in hand to develop automated systems for monitoring facial expressions and animating artificial models. Face technology of the sort we describe, which is now feasible and achievable within a relatively short time frame, could revolutionize fields as diverse as medicine, law, communications, and education.

In this report we summarize the scientific and engineering issues that arise in meeting those challenges and outline recommendations for achieving those goals. First we provide background on the importance of facial expression.

I-B. Background About the Importance of Facial Expression

Facial expressions provide information about:

- affective state, including both emotions such as fear, anger, enjoyment, surprise, sadness, disgust, and more enduring moods such as euphoria, dysphoria, or irritableness;
- cognitive activity, such as perplexity, concentration, or boredom;
- temperament and personality, including such traits as hostility, sociability or shyness;
- truthfulness, including the leakage of concealed emotions, and clues as to when the information provided in words about plans or actions is false;
- psychopathology, including not only diagnostic information relevant to depression, mania, schizophrenia, and other less severe disorders, but also information relevant to monitoring response to treatment.

In basic research on the brain, facial expressions can identify when specific mental processes are occurring, periods that can be examined with new imaging technologies (e.g., Magnetic Resonance Imaging or Positron Emission Tomography) too expensive for continuous monitoring. Facial expressions also hold promise for applied medical research, for example, in identifying the role of emotions and moods in coronary artery disease.

In education, the teacher's facial expressions influence whether the pupils learn and the pupil's facial expressions can inform the teacher of the need to adjust the instructional message.

In criminal justice contexts, facial expressions play a crucial role in establishing

or detracting from credibility.

In business, facial expressions are important in negotiations and personnel decisions.

In medicine, facial expressions can be useful in studies of the autonomic nervous system and the psychological state of the patient.

In international relations, analysis of facial expressions of world leaders can be used for assessing reliability and change in emotional state.

In man-machine interactions, facial expressions could provide a way to communicate basic information about needs and demands to computers.

I-C. Progress and Obstacles in Measuring Facial Expression

Development of methods for measuring the movement of the facial muscles, which produce the facial expressions, has taken great strides recently. Research using these new methods has shown promise, uncovering many findings in the diverse areas listed above. There now is an international community of more than two hundred facial researchers, but their progress is slowed because the facial measurement methods available are very labor intensive -- requiring more than one hour to measure each minute of facial expression.

I-D. The Opportunity

Recent developments in computer vision and neural networks, used primarily in lip reading, and the recognition of specific individuals from static faces, indicate the possibility of being able to automate facial measurement. Work is underway at a number of laboratories both in the United States and abroad, in university settings and in businesses, to take the next steps in developing semi or fully automated procedures for extracting information from faces.

The National Science Foundation has a crucial role to play in maintaining the U.S. technological lead in this area, in light of well funded efforts in other countries. There is more work to be done than can be funded only by NSF and other federal agencies. Foundations and the private sector will need to be involved.

I-E. Recommendations

This report contains a list of the major unresolved research issues and needed technological innovations related to efforts to measure the face automatically. Many basic science questions on how to interpret the messages of the face, how messages function in communication, and how they are produced and interpreted by the neuro-muscular system, remain open. Answering these questions is prerequisite for developing and applying automated measurements fully, but answering these question would, in turn, be expedited by automated measurement tools. Development of these tools is an interdisciplinary project requiring input from many areas of engineering, computer science,

neuroscience, and behavioral science. Such efforts require both an infrastructure, such as multimedia databases and image processing computers, and training and education, such as post-doctoral programs and electronic communications, to support cooperation among investigators. Recommended tools, infrastructure elements, and educational resources are also listed below in the Recommendations, pages 53 to 57.

I-F. Organization of the Report

First the origin, planning, and goals of the Workshop are described. Then the Workshop's schedule and participants are listed.

Following this overview of the Workshop is a summary of each of the tutorial sessions. (The separate tutorials on Neural Networks and on Finding, Organizing and Interpreting Faces have been combined into one report).

Next are reports from each of the breakout work groups assigned to plan the steps for progress in areas relevant to this endeavor: basic science, sensing and processing, and databases and modeling.

A list of recommendations, integrated from across the different sections, follows. A discussion of the benefits of research and development in this field is in the final section.

II. OVERVIEW OF THE WORKSHOP

Paul Ekman

II-A. Organization of the Workshop

A Workshop on Facial Understanding was held in response to a request from NSF to determine the research problems in the area of how to extract information from the face using new developments in such areas as computer science, neural networks, and image processing. The organizing committee consisted of Paul Ekman, University of California at San Francisco, a psychologist who has done extensive work on facial expression; Tom Huang, University of Illinois, an electrical engineer who has worked extensively in computer vision; and Terry Sejnowski, University of California at San Diego and Salk Institute, a neuroscientist who has applied neural networks to a variety of problems in visual pattern recognition. The scope and organization of the Workshop were decided by this organizing committee.

II-B. Scope of the Workshop

It was necessary to bring the diverse participants up to date on relevant developments, in fields other than their own, that are relevant to the focus of the Workshop. Tutorial plenary sessions addressed: (a) Psychology and neuroanatomy of facial expression; (b) Computer vision, motion processing and lip reading; (c) Neural networks; (d) Finding, organizing and interpreting faces; and (e) Hardware considerations.

Breakout work groups dealt with three general themes: (a) Basic science (what about the face is represented; what type of information from facial behavior is abstracted and quantified); (b) Sensing and processing (the preliminary analysis transformations, recording methods, sensors, preprocessing, motion analysis); and (c) Modeling and data structure (organization of a facial database; access methods, type of information stored; representation or display of visual facial information; how to use information in the data base; modeling, recreation of faces, effecter systems, robotics).

Each breakout work group identified the important research issues and the infrastructure that is needed to facilitate research.

II-C. Goals achieved

The Workshop:

- identified the most important areas of research on how to extract information from facial activity relevant to a person's emotional, cognitive and physical state.

- considered computer vision techniques for facial processing and categorization of facial expressions relevant to enhancing communication between man and machine.
- considered how we can facilitate the training of new investigators in the relevant fields; including what curricula is needed in the education of new investigators, and what tools, communication media, support structures, etc. are needed to support basic research over the next several years.

II-D. Workshop Schedule

The Workshop was held in Washington D.C. July 30-August 1st. The main sections of the Workshop are listed below.

Day 1:
Participants gave five minute summaries of their own work
Tutorial Session 1: Psychology and Neuroanatomy of Facial Expression
Tutorial Session 2: Computer Vision, Motion Processing and Lip reading.
Tutorial Session 3: Neural Networks
Tutorial Session 4: Finding, organizing and interpreting faces.

Day 2:
Tutorial Session 5: Hardware Considerations
Breakout Workgroup Meetings
Reports by Breakout groups

Day 3:
Breakout group meetings to develop recommendations
Report from Breakout groups

II-E. Participants

The selection of participants was based on (a) the concerns of the organizing committee to represent different areas of expertise within the broad areas relevant to the Workshop; (b) recommendations from NSF; (c) the desire to have representatives from universities, government research labs, and industry research labs. To facilitate discussion we decided to limit invitations to under 30.

Name	Position	Institution
Y. T. Chien	Division Director Division of Information, Robotics and Intelligent Systems (IRIS)	National Science Foundation
Howard Moraff	Program Director Robotics and Machine Intelligence Program	National Science Foundation
Su-Shing Chen	Program Director Knowledge Models and Cognitive Systems Program	National Science Foundation
John Hestenes	Program Director Interactive Systems Program	National Science Foundation
Richard Louttit	Acting Division Director Division of Behavioral and Cognitive Science	National Science Foundation
Jean		National Science

Jean Intermaggio	Program Director Social Psychology Program	National Science Foundation
Rodney Cocking	Chief, Cognition, Learning and Memory	National Institute of Mental Health
Lynn Huffman	Chief, Personality & Emotion Program	National Institute of Mental Health
Richard Nakamura	Chief, Cognition & Behavioral Neuroscience	National Institute of Mental Health
Wil Irwin	Human Interaction Lab	UC San Francisco
Ron Cole	Center for Spoken Language Understanding	Oregon Graduate Institute
Beatrice Golomb	University of California Medical School	UCLA

Name	Unit	Institution	City/State/Country
John Allman	Division of Biology 216-76	California Institute of Technology	Pasadena, CA 91125
Dana Ballard	Department of Computer Science	University of Rochester	Rochester, NY 14627
Gordon Baylis	Department of Psychology	University of California, San Diego	La Jolla, CA 92093
Peter Burt	CN 5300	Sarnoff Research Center	Princeton, NJ 08543
John T. Cacioppo	Dept. of Psychology	Ohio State University	Columbus, OH 43210-1222
Gary Cottrell	Department of Computer Science and Engineering	University of California, San Diego	La Jolla, CA 92093
Richard J. Davidson	Psychophysiology Laboratory	University of Wisconsin	Madison, WI 53706
Paul Ekman	Human Interaction Lab	University of California	San Francisco, CA 94143
Wallace V. Friesen		234 Hillsboro	Lexington, KY 40511
Oscar Garcia	Dept. of Electrical Engineering and Computer Science	The George Washington University	Washington DC 20052
Joseph C. Hager	Human Interaction Lab	University of California	San Francisco, CA 94143
Thomas S. Huang	Coordinated Science Laboratory	University of Illinois	Urbana, IL 61801

Carroll Izard	Dept. of Psychology	University of Delaware	Newark, DE 19716
Hsien-Che Lee	Imaging Research Labs	Eastman Kodak Company	Rochester NY 14650-1816
Kenji Mase	NIT R&D Information & Patent Center	Nippon Telegraph & Telephone Corporation	1-1-7, Uchisaiwai-cho, Chiyoda-ku, Tokyo 100 JAPAN
Fred Parke	PS/LOB Technology, 9462	IBM	Austin, TX 78758
Alexander Pentland	Rm E-15-187 The Media Lab	MIT	Cambridge, MA 02139
Michael Phillips	MIT Computer Science Lab	MIT/TD>	Cambridge MA 02139
Demetri Psaltis	Department of Electrical Engineering	California Institute of Technology	Pasadena, CA 91125
Louis Sadler	Head, Dept. of Biomedical Visualization	University of Illinois	Chicago, IL 60612 (M/C 527)
Terrence J. Sejnowski	Computational Neurobiology Lab	Salk Institute	San Diego, CA 92186-5800
Peter Stucki	Institute for Informatics (Dept. of CS)	Univ. of Zurich	Zurich Switzerland
Demetri Terzopoulos	Dept. of CS	University of Toronto	Toronto, CANADA M5S 1A4
Thomas Vetter	MIT AI Lab	MIT	Cambridge, MA 02139
Lance Williams		Apple Computer, Inc.	Cupertino, CA 95014
Alan Yuille	Division of Applied Physics	Harvard University	Cambridge, MA 02138
Jun Zhang		Salk Institute	San Diego, CA 92186

III. TUTORIAL SUMMARIES

III-A. The Psychology and Neuroanatomy of Facial Expression

John Cacioppo, Joseph Hager, and Paul Ekman

Abstract. This session surveys the different sources of information in the face and the different types of information that can be derived. Neural efferent pathways include the brain areas transmitting to the facial nerve, the facial nerve to facial nucleus, and the facial nucleus to muscles. The relationship between electromyographic (EMG) measurement, muscle tonus measurements, and visible observable facial activity and methods is considered. Evidence on emotion signals includes universals, development, spontaneous versus deliberate actions, and masked emotions. The face also provides conversational signals and signs relevant to cognitive activity. The logic of comprehensively measuring facial movement illustrates how FACS scores facial behavior, the mechanics of facial movement, and options for what to score (intensity, timing, symmetry). Relationships between facial behavior, voice, and physiological measures are discussed. A database of the face and support for implementing this resource are needed.

Presenters: J. T. Cacioppo, P. Ekman, W. V. Friesen, J. C. Hager, C. E. Izard

Facial Signal Systems

The face is the site for the major sensory inputs and the major communicative outputs. It is a multisignal, multimessage response system capable of tremendous flexibility and specificity (Ekman, 1979; Ekman & Friesen, 1975). This system conveys information via four general classes of signals or sign vehicles: (1) *static facial signals* represent relatively permanent features of the face, such as the bony structure and soft tissues masses, that contribute to an individual's appearance; (2) *slow facial signals* represent changes in the appearance of the face that occur gradually over time, such as the development of permanent wrinkles and changes in skin texture; (3) *artificial signals* represent exogenously determined features of the face, such as eyeglasses and cosmetics; and (4) *rapid facial signals* represent phasic changes in neuromuscular activity that may lead to visually detectable changes in facial appearance. (See Ekman, 1978, for discussion of these four signal systems and eighteen different messages that can be derived from these signals).

All four classes of signals contribute to facial recognition. We are concerned here, however, with rapid signals. These movements of the facial muscles pull the skin, temporarily distorting the shape of the eyes, brows, and lips, and the

appearance of folds, furrows and bulges in different patches of skin. These changes in facial muscular activity typically are brief, lasting a few seconds; rarely do they endure more than five seconds or less than 250 ms. The most useful terminology for describing or measuring facial actions refers to the production system -- the activity of specific muscles. These muscles may be designated by their Latin names, or a numeric system for Action Units (AUs), as is used in Ekman and Friesen's Facial Action Coding System (FACS, see page 10). A coarser level of description involves terms such as smile, smirk, frown, sneer, etc. which are imprecise, ignoring differences between a variety of different muscular actions to which they may refer, and mixing description with inferences about meaning or the message which they may convey.

Among the types of messages conveyed by rapid facial signals are: (1) emotions -- including happiness, sadness, anger, disgust, surprise, and fear; (2) emblems -- culture-specific symbolic communicators such as the wink; (3) manipulators -- self-manipulative associated movements such as lip biting; (4) illustrators -- actions accompanying and highlighting speech such as a raised brow; and (5) regulators -- nonverbal conversational mediators such as nods or smiles (Ekman & Friesen, 1969).

A further distinction can be drawn among rapid facial actions that reflect: (1) reflex actions under the control of afferent input; (2) rudimentary reflex-like or impulsive actions accompanying emotion and less differentiated information processing (e.g., the orienting or defense response) that appear to be controlled by innate motor programs; (3) adaptable, versatile, and more culturally variable spontaneous actions that appear to be mediated by learned motor programs; and (4) malleable voluntary actions. Thus, some classes of rapid facial actions are relatively undemanding of a person's limited information processing capacity, free of deliberate control for their evocation, and associated with (though not necessary for) rudimentary emotional and symbolic processing, whereas others are demanding of processing capacity, are under voluntary control, and are governed by complex and culturally specific prescriptions, or display rules (Ekman & Friesen, 1969), for facial communications. (The terms facial "actions," "movements," and "expressions" are used interchangeably throughout this report).

Techniques for Measuring the Rapid Facial Signals

Numerous methods exist for measuring facial movements resulting from the action of muscles (see a review of 14 such techniques in Ekman, 1982; also Hager, 1985 for a comparison of the two most commonly used, FACS and MAX). The Facial Action Coding System (FACS) (Ekman and Friesen, 1978) is the most comprehensive, widely used, and versatile system. Because it is being used by most of the participants in the Workshop who currently are working with facial movement, and is referred to many times in the rest of this report, more detail will be given here about its derivation and use than about other techniques. Later, the section on neuroanatomy of facial movement (page 12) discusses electromyography (EMG), which can measure activity that might not be visible, and, therefore, is not a social signal.

The Facial Action Coding System (FACS)

FACS was developed by determining how the contraction of each facial muscle (singly and in combination with other muscles) changes the appearance of the face. Videotapes of more than 5000 different combinations of muscular actions were examined to determine the specific changes in appearance which occurred and how to best differentiate one from another. It was not possible to reliably distinguish which specific muscle had acted to produce the lowering of the eyebrow and the drawing of the eyebrows together, and therefore the three muscles involved in these changes in appearance were combined into one specific Action Unit (AU). Likewise, the muscles involved in opening the lips have also been combined.

Measurement with FACS is done in terms of Action Units rather than muscular units for two reasons. First, for a few changes in appearance, more than one muscle has been combined into a single AU, as described above. Second, FACS separates into two AUs the activity of the frontalis muscle, because the inner and outer portion of this muscle can act independently, producing different changes in appearance. There are 46 AUs which account for changes in facial expression, and 12 AUs which more grossly describe changes in gaze direction and head orientation.

Coders spend approximately 100 hours learning FACS. Self instructional materials teach the anatomy of facial activity, i.e., how muscles singly and in combination change the appearance of the face. Prior to using FACS, all learners are required to score a videotaped test (provided by Ekman), to insure they are measuring facial behavior in agreement with prior learners. To date, more than 300 people have achieved high inter-coder agreement on this test.

A FACS coder "dissects" an observed expression, decomposing it into the specific AUs which produced the movement. The coder repeatedly views records of behavior in slowed and stopped motion to determine which AU or combination of AUs best account for the observed changes. The scores for a facial expression consist of the list of AUs which produced it. The precise duration of each action also is determined, and the intensity of each muscular action and any bilateral asymmetry is rated. In the most elaborate use of FACS, the coder determines the onset (first evidence) of each AU, when the action reaches an apex (asymptote), the end of the apex period when it begins to decline, and when it disappears from the face completely (offset). These time measurements are usually much more costly to obtain than the decision about which AU(s) produced the movement, and in most research only onset and offset have been measured.

The FACS scoring units are descriptive, involving no inferences about emotions. For example, the scores for a upper face expression might be that the inner corners of the eyebrows are pulled up (AU 1) and together (AU 4), rather than that the eyebrows' position shows sadness. Data analyses can be done on these purely descriptive AU scores, or FACS scores can be converted by a computer using a dictionary and rules into emotion scores. Although this emotion interpretation dictionary was originally based on theory, there is now considerable empirical support for the facial action patterns listed in it:

- FACS scores yield highly accurate pre- and postdictions of the emotions signaled to observers in more than fifteen cultures, Western and non-Western, literate and preliterate (Ekman, 1989);
- specific AU scores show moderate to high correlations with subjective reports by the expresser about the quality and intensity of the felt emotion (e.g., Davidson et al., 1990);
- experimental circumstances are associated with specific facial expressions (Ekman, 1984);
- different and specific patterns of physiological activity co-occur with specific facial expressions (Davidson et al. 1990).

The emotion prediction dictionary provides scores on the frequency of the seven single emotions (anger, fear, disgust, sadness, happiness, contempt, and surprise), the co-occurrence of two or more of these emotions in blends, and a distinction between emotional and nonemotional smiling, which is based on whether or not the muscle that orbits the eye (AU 6) is present with the muscle that pulls the lip corners up obliquely (AU 12). Emotional smiles are presumed to be involuntary and to be associated with the subjective experience of happiness and associated physiological changes. Nonemotional smiles are presumed to be voluntary, and not to be associated with happy feelings nor with physiological changes unique to happiness. A number of lines of evidence -- from physiological correlates to subjective feelings -- now support this distinction between emotional and nonemotional smiles (reviewed in Ekman, 1992a).

The Maximally Discriminative Affect Coding System (MAX)

Izard's (1979) MAX also measures visible appearance changes in the face. MAX's units are formulated in terms of appearances that are relevant to eight specific emotions, rather than in terms of individual muscles. Unlike FACS, MAX does not exhaustively measure all facial actions, but scores only those facial movements Izard relates to one or more of the eight emotions. All of the facial actions which MAX specifies as relevant to particular emotions are also found in the FACS emotion dictionary, but that database contains inferences about many other facial actions not present in MAX that may signal emotion. There is some argument (Oster et al., in press) about whether the facial actions MAX specifies as relevant to emotion are valid for infants.

Evidence About Which Facial Actions Signal Which Emotions

The scientific study of what facial configurations are associated with each emotion has primarily focused on observers' interpretations of facial expressions (e.g., judgements of pictures of facial expressions). There has been far less research, although some, that has examined how facial expressions relate to other responses the person may emit (i.e., physiological activity, voice, and speech) and to the occasion when the expression occurs. (1) Across cultures there is highly significant agreement among observers in categorizing facial expressions of happiness, sadness, surprise, anger, disgust, and fear. (Note there have been some recent challenges to this work, but they are ideologically, not empirically, based and are proposed by those who claim emotions do not exist [Fridlund, 1991] or that emotions are socially constructed

and have no biological basis [Russell, 1991a,b,c]). (2) The experimental inductions of what individuals report as being positive and negative emotional states are associated with distinct facial actions, as are the reports of specific positive and specific negative emotions. (3) Cultural influences can, but do not necessarily, alter these outcomes significantly. (4) These outcomes can be found in neonates and the blind as well as sighted adults, although the evidence on the blind and neonates is more limited than that for sighted adults. (5) Emotion-specific activity in the autonomic nervous system appears to emerge when facial prototypes of emotion are produced on request, muscle by muscle. (6) Different patterns of regional brain activity coincide with different facial expressions. (7) The variability in emotional expressions observed across individuals and cultures is attributable to factors such as differences in which emotion, or sequence of emotions, was evoked and to cultural prescriptions regarding the display of emotions (e.g., Ekman, 1972, 1992b; Ekman & Friesen, 1978; Ekman et al., 1972, 1983; Izard, 1971, 1977).

Facial actions have also been linked to nonemotional information processing. For instance, in addition to nonverbal messages (e.g., illustrators or emblems; see Ekman & Friesen, 1969; Ekman, 1979), incipient perioral activity has been observed during silent language processing (Cacioppo & Petty, 1981; McGuigan, 1970), increased activity over the eyebrow region (corrugator supercilii) and decreased blinking has been associated with mental concentration or effort (e.g., Darwin, 1872; Cacioppo et al., 1985; Stern & Dunham, 1990), and gesture/speech mismatches have been observed during the simultaneous activation of incompatible beliefs (Goldin-Meadow et al., in press).

The Neuroanatomy of Facial Movement

Rapid facial signals, such as emotional expressions, are the result of movements of facial skin and connective tissue (i.e., fascia) caused by the contraction of one or more of the 44 bilaterally symmetrical facial muscles. These striated muscles fall into two groups: four of these muscles, innervated by the trigeminal (5th cranial) nerve, are attached to and move skeletal structures (e.g., the jaw) in mastication; and forty of these muscles, innervated by the facial (7th cranial) nerve, are attached to bone, facial skin, or fascia and do not operate directly by moving skeletal structures but rather arrange facial features in meaningful configurations (Rinn, 1984). Although muscle activation must occur if these facial configurations are to be achieved, it is possible for muscle activation to occur in the absence of any overt facial action if the activation is weak or transient or if the overt response is aborted.

Briefly, the neural activation of the striated muscles results in the release of acetylcholine at motor end plates, which in turn leads to muscle action potentials (MAPs) that are propagated bidirectionally across muscle fibers and activate the physiochemical mechanism responsible for muscle contraction. The activating neurotransmitter acetylcholine is quickly eradicated by the enzyme acetyl cholinesterase, so that continued efferent discharges are required for continued propagation of MAPs and fiber contraction. Moreover, low amplitude neural volleys along motor nerves tend to activate small motoneurons, which innervate relatively few and small muscle fibers (a relationship called the size

principle; Henneman, 1980). Thus, dynamic as well as configural information flows from the muscles underlying rapid facial signals (Cacioppo & Dorfman, 1987). Finally, fast or low-level changes in these efferent discharges can occur without leading to one-to-one feature distortions on the surface of the face. This is due to factors such as the organization of the facial muscles (e.g., agonist/antagonist, synergist, and how some muscles overlay other muscles) and the structure and elasticity of the facial skin, facial sheath, adipose tissue, and facial muscles. Not unlike a loose chain, the facial muscles can be pulled a small distance before exerting a significant force on the object to which they are anchored (Ekman, 1982; Tassinary et al., 1989). In addition, the elasticity of the facial sheath, facial skin, and adipose tissue acts like a low-pass mechanical filter. Therefore, electromyography (measuring electrical activity by attaching electrodes to the surface of the face), has served as a useful complement to overt facial action coding systems (see review by Cacioppo et al., 1990).

The muscles of mimicry on each side of the face are innervated by a lower motor nerve emanating from a facial nerve nucleus located in the pons. The left and right facial nerve nuclei are independent, but the commands they carry to the lower face are from contralateral upper motorneuron tracts whereas the instructions they carry to the mid-upper face are from bilateral motorneuron tracts. The upper motorneuron tracts include corticobulbar and subcortical ("extrapyramidal") pathways. Lesions in the former are associated with hemiparalysis of voluntary movements, whereas lesions in the latter are more typically associated with attenuated spontaneous movements (Rinn, 1984). Despite these distinctions, the subcortical and cortical control mechanisms provide complementary influences, with (1) the former well suited for spontaneous, non-flexible behaviors that are directly and immediately in the service of basic drives and (2) the latter providing adaptability by allowing learning and voluntary control to influence motor behavior.

Facial Data Base

A readily accessible, multimedia database shared by the diverse facial research community would be an important resource for the resolution and extension of issues concerning facial understanding. This database should contain images of faces (still and motion), vocalizations and speech, research findings, psychophysiological correlates of specific facial actions, and interpretations of facial scores in terms of emotional state, cognitive process, and other internal processes. The database should be supplemented with tools for working with faces, such as translating one facial measurement system into another, and modifying the expression of an image synthetically. Such a database would be capable of facilitating and integrating the efforts of researchers, highlighting contradictions and consistencies, and suggesting fruitful avenues for new research. Only isolated pieces of such a database exist now, such as the FACS Dictionary (Friesen and Ekman, 1987), which needs updating.

Note: This report was originally prepared by John Cacioppo and Joseph Hager, and then revised and edited by Paul Ekman. It is based on presentations by each of these authors, C. E. Izard, and W. V. Friesen.

III-B. Computer Vision and Face Processing

Thomas S. Huang, Peter Burt, and Kenji Mase

Abstract. Computer vision deals with the problem of scene analysis; more specifically the extraction of 3D information about scenes/objects from 2D (possibly time-varying) images obtained by sensors such as television cameras. Over the years, many algorithms have been developed for determining 3D shape, texture, and motion. In this session, some of the major approaches are reviewed in light of applications to face structure/motion analysis. Emphasis is on methods of estimating 3D rigid and nonrigid motion/structure from 2D image sequences.

Presenters: T. S. Huang, P. Burt, K. Mase

Introduction

In this brief tutorial, we review computer vision with special emphasis on face processing.

From an engineering viewpoint, the goal of computer vision is to build automated systems capable of analyzing and understanding 3D and possibly time-varying scenes. The input to such systems are typically two-dimensional (2D) images taken, for example, by television cameras, although sometimes direct range sensors are also used. There are three levels of tasks in computer vision systems. (i) Level 1 (lowest): Reconstruction -- to recover 3D information, both geometrical (shape) and photometric (texture), from the sensed data. (ii) Level 2 (middle): Recognition -- to detect and identify objects. (iii) Level 3 (highest): Understanding -- to figure out what is going on in the (possibly time-varying) scene.

Representation

Representation is a central issue in object recognition and scene understanding. The desirable features of a representation are: It should be easy to construct (from sensed data), easy to update, easy to use (for particular applications such as object recognition), and efficient to store. There are two main classes of 3D shape representation methods: surface and volume (Chen & Huang, 1988; Agin & Binford, 1973; Requicha, 1980). The spatial relationships between objects can be represented by relational graphs (structures) (Barrow & Popplestone, 1971).

Most work on object representation in robot vision has been on the geometrical aspects. Relatively little has been done on the photometric aspects of object representation.

Reconstruction

By reconstruction, we mean the determination of the 3D coordinates of points on the object surface and some of the reflectance properties. Again, most work in computer vision has been on the geometrical aspects.

The use of laser range finders is the most direct way of obtaining ranges of points on an object surface. A laser beam hits a surface point, and the reflected light is caught by a sensor. The range of the point is determined by either the time-of-flight or the phase shift of the beam. By scanning the laser beam, one can obtain a range map. A recent example is the ERIM laser scanner which uses the phase-shift method. It gives a 256 x 256 range map as well as several bands of intensity images in about 2 seconds. The maximum range is about 32 feet, and the accuracy about plus or minus 0.5 feet.

The 3D shape of an object surface 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 (Will & Pennington, 1971). For example, if a square grid is projected on a plane (in 3D), then the images of the grid lines are still straight lines, and from the skewness of the grid, one can determine the orientation of the plane.

Another method is laser illuminated triangulation. The method involves a laser and a camera. The geometry of the setup is known. The laser beam illuminates a spot on the object surface. An image of the surface is taken, and the 2D coordinates of the image of the spot are measured. Then, by triangulation, the 3D coordinates of the point are determined. By changing the direction of the laser beam, one can obtain 3D coordinates of different points on the surface.

The classical technique of photogrammetry (Moffitt & Mikhail, 1980) is passive stereo. While the methods described above are "active" in the sense that the illumination is controlled, this method is "passive" -- the illumination is natural. Two images of an object are taken from different viewpoints. Then: (i) corresponding points are found between the two images. (ii) for each corresponding pair, the 3D coordinates of the point are determined by triangulation. The first step, that of finding correspondences, is extremely difficult. Typically, for each point in one of the images, one aims to find the corresponding point in the other image. This is usually done by some sort of cross-correlation. In computer vision, seminal work in stereo was done by Marr (1982).

Usually, it is not possible to find point correspondences in some regions of the images (e.g., where the intensity is almost uniform). Thus, 3D surface fitting (interpolation) is necessary (Grimson, 1983).

The most difficult problem in reconstruction is to extract 3D information of an object from a single 2D image. The pioneering work is that of Horn (1986), who investigated the problem of shape from shading, i.e., to extract surface orientation information of an object from a 2D intensity image of it. Equations can be written, relating the normal direction at a 3D surface point and the observed intensity of the corresponding image point, which involve the reflectance function at the 3D point and the illumination parameters. The

difficulty is that there are too many unknown variables, making the problem basically underdetermined.

Witkin (1981) was the first to investigate the recovery of surface orientation from texture. If we assume that the surface texture is actually isotropic in 3D, then from the anisotropicity of the observed image texture, the surface orientation in 3D may be deduced. For example, we can estimate the orientation of a planar lawn relative to the camera geometry from the local density variation of the grass in the image.

Kanade (1981) was the first to investigate the recovery of surface orientation from 2D shapes in the image. If we can assume that the 3D object has certain symmetry, then surface orientation may be recovered from the skew symmetry of the image. For example, if we can assume that an observed 2D ellipse is the perspective view of a circle in 3D, then the orientation of the circle may be determined.

More recently, interesting work has been done using color (Lee, 1986) and polarization (Wolff, 1988).

Object Recognition

The problem is to recognize a 3D object from one (or sometimes several) 2D views of it. By recognition, one usually means to classify the object into one of a set of prescribed classes. In an easy case, all objects belonging to the same class look almost exactly the same, and the number of classes is small. In a harder case, the number of classes is very large -- e.g., in fingerprint identification. In the hardest case, the classes are generic, i.e., objects in the same class may look very different -- e.g., to recognize a "chair."

The main approach to recognition of a 3D object from a single 2D view is as follows. We have a 3D model for each class. The observed 2D view of the unknown object is compared with many 2D views generated from the 3D models of candidate classes. The best match determines the class. In practice, a major problem is how to reduce the search space (the number of candidate classes, and the number of 2D views from each 3D model) by using *a priori* information, heuristics, etc. For example, in vehicle classification, to detect and count the number of wheels can help to limit the number of candidate classes. Thus, if there are more than two wheels in the side view of a vehicle, it cannot be a sedan (Leung & Huang, 1992).

A general way of reducing the number of 2D views of a 3D model (which one has to generate and compare with the 2D view of the unknown object) is the concept of aspect graph (Koenderink & van Doorn, 1979; Eggert & Bowyer, 1989; Ponce & Kriegman, 1989). An aspect graph enumerates all possible "qualitative" aspects an object may assume. We partition the viewing space into regions such that viewpoints within the same region give 2D views of the object which are "qualitatively" similar. A change in "quality" occurs only when the viewpoint crosses a region boundary. Thus, it is necessary to compare the 2D view of the unknown object with only one typical 2D view from each region. Of course, for a given application one has to define precisely what is meant by

"quality." And the definition of quality will obviously influence the matching criterion. E.g., if the 3D object is a polyhedron, then the "quality" of a 2D view could be the numbers of faces, edges, and vertices.

Motion Analysis and Face Processing

The role of motion in recognizing facial expressions

Automated analysis of image motion can play several roles in monitoring and recognizing human facial expressions. A motion system is needed to detect and track the head and face. Then, within the face, a motion system is needed to track motions or deformations of the mouth, eyes, and other facial structures involved in the generation of expressions. While expressions can often be recognized from a single snapshot of a face, the ability to discriminate subtle expressions requires a comparison over time as the face changes shape. In addition, the temporal pattern of changes in facial expression observed through motion analysis may carry important information. Particularly rapid and precise motion analysis of lip motion is needed for lip reading. At the other extreme, a motion analysis system is needed to monitor the gross motions of the head, body and hands to provide gesture recognition for next generation computer interfaces (Huang & Orchard, 1992).

Three approaches to motion analysis

The approaches to motion analysis that have been developed in computer vision fall broadly into three classes (Huang, 1987). Each has advantages and disadvantages for applications to the recognition of facial expressions.

Feature Tracking:

In the feature tracking approach to motion analysis, motion estimates are obtained only for a selected set of prominent features in the scene. Analysis is performed in two steps: first each image frame of a video sequence is processed to detect prominent features, such as edges or corner-like patterns, then the features are matched between frames to determine their motion. An advantage of this approach is that it achieves efficiency by greatly reducing image data prior to motion analysis, from a quarter million pixels to a few hundred features. A disadvantage is that motion vectors are not obtained for all points in the scene. When applied to a face it may not find vectors for all parts of the face that contribute to an expression.

Flow:

In the flow approach to motion analysis, motion vectors are estimated at a regular array of points over the scene. The motion vector at a given point can be based on local gradients in space and time at that point, or on the cross correlation of the pattern in the neighborhood of the point between successive frames. Advantages of the flow approach are that a dense array of motion vectors is obtained, and that processing is well suited for special purpose hardware. Disadvantages are that local estimates of flow tend to be "noise" and that prodigious computations are required.

Pattern Tracking:

Pattern tracking is similar to feature tracking in that patterns are first located in each image frame, then these are followed from frame to frame to estimate motion. It differs from feature tracking in that "high level" patterns are used that are indicative of the objects of interest, rather than "low level" generic features, such as edge elements. In the case of face tracking, a pattern-based method might search for an entire face, using a set of face templates to span a range of face types, orientations, and scales, or it might search first for distinctive parts of the face, such as the eyes. Efficient techniques restrict the search to regions of the scene where a face is likely to occur based on where it was found in the previous frame, and reduce the resolution and sample density of the source images to be the minimum sufficient for the task. Advantages of the pattern tracking approach are that it is directed to just the objects of interest in the scene, such as faces; motion is not computed for extraneous background patterns. Disadvantages are that an enormous number of relatively complex patterns may need to be searched to identify faces over a wide range of viewing conditions.

Model based alignment

Advanced methods for motion analysis make use of models to constrain the estimation process. Models range from simple and generic to complex and object specific. In the first category are "smoothness" constraints that look for solutions to the motion equations that vary smoothly everywhere in a scene, except at object boundaries. These constraints are implicit in almost all approaches to motion analysis and reflect the fact that physical objects are locally rigid. In the second category are object model constraints that, in effect, align a 3D model of an object, such as a face and head, to the image data in order to estimate both object motion and orientation (pose).

The use of models can significantly improve the efficiency and precision of motion estimates since analysis considers only the motions that are physically possible. At the same time it facilitates image interpretation by analyzing motion into physically meaningful components. For example, observed motions of a face can be separated into the motions of the head within the camera-centered coordinate system, and the motions of parts of the face within a head-centered coordinate system. This separation is essential for detecting subtle facial expressions.

State of technology

The analysis of image motion is perhaps the most active subarea of research in computer vision. Methods that provide sufficient detail and precision for recognizing facial expression and that can be computed at video rates are only now becoming feasible. Hardware capable of real time motion analysis is being built for specialized tasks, including video compression and moving target detection. Still, significant further development is required before "off-the-shelf" technology will be available for application to recognizing facial expressions.

Optic Flow Approach to Face Processing

Once head position is located by any means, we can examine the movements of facial actions. The human face has several distinctive features such as eyes, mouth and nose. Facial skin at cheek and forehead has the texture of a fine-grained organ. These features and texture are exploitable as clues in extracting and then recognizing facial expressions by computer vision techniques.

In the following, we give examples of applying existing computer vision techniques and experimental results to expression recognition and lip reading.

Extraction of facial movement

Muscle actions can be directly observed in image sequence as optical flow, which is calculated by facial features and skin deformation. A gradient based optical flow algorithm has better characteristics for face analysis than other algorithms such as correlation based, filtering based, and token-matching based. For instance, it does not require feature extraction and tracking processes but only assumes smooth, short range translational motion. Dense flow information was computed from image sequences of facial expression with Horn and Schunck's (1981) basic gradient algorithm. The computed flow can capture skin deformation fairly well, except for the area where the aperture problem and the motion discontinuity problem arise. The aperture problem refers to the fact that if a straight edge is in motion locally it is possible to estimate only the component of the velocity that is orthogonal to the edge. Abrupt changes of the flow field cause problems because the Horn-Schunck algorithm uses the smoothness constraint.

Extracting muscle movement

The dense optical flow information to each muscle (group) action is reduced by taking the average length of directional components in the major directions of muscle contraction. Several muscle windows are located manually to define each muscle group using feature points as references. The muscle windows are regarded as a muscle (group) model which has a single orientation of muscle contraction/relaxation. Experiments assured that some important action could be extracted and be illustrated as functions of time.

The muscle action derived from the muscle (group) model can be associated with several Action Units (AUs) of the Facial Action Coding System (Ekman & Friesen, 1978).

Thus, for example, an estimated muscle motion of a happy expression could be scored as AUs 6(0.6) + 12(0.35) + 17(0.28) + 25(0.8) + 26(0.20). The figures in the parenthesis, e.g. 0.6 for AU 6, indicate the AU's strength, which is equated to the maximum absolute velocity (pixels/frame) within a short sequence.

Recognition of facial expression

Since optical flow of a facial image sequence contains rich information of facial actions of various expressions, conventional pattern classification techniques as well as neural networks can be used for recognition of expressions. As a

preliminary experiment, a feature vector whose elements are the means and the variances of optical flow data at evenly divided small regions (blocks) is used. The dimensionality is reduced to make the vector concise by introducing a separation goodness criterion function, which is similar to Fisher's criterion (Mase, 1991).

In the recognition experiment of four expressions in motion, e.g. happiness, anger, surprise and disgust, 19 out of 22 test data were correctly identified (Mase & Pentland, 1990a).

Lipreading

Approaches similar to those used for facial expression recognition can be used for lip reading (Duda & Hart, 1973). Four muscle windows are located around the mouth; above, below, left, and right, to extract optical flow data related to speech generation. The mean values of each flow vector component within a window are computed, as we did in expression recognition. The principle component analysis on the training set of English digits produces two principle parameters, i.e., mouth open and elongation, which are also acceptable intuitively. The experiments on the continuously spoken data by four speakers shows over 70% accuracy of digit recognition including word segmentation. Since optical flow becomes zero at motion stop and reverse, the approach has good advantage in temporal segmentation of facial expression as well as lipreading.

Note: This report was prepared by T. S. Huang, P. Burt, and K. Mase, based on Workshop presentations by these authors, and edited by T. S. Huang.

III-C. Neural Networks and Eigenfaces for Finding and Analyzing Faces

Alexander Pentland and Terrence Sejnowski

Abstract: Animal communication through facial expressions is often thought to require very high-level, almost cognitive processing, yet pigeons can be trained to recognize faces and babies can do so as early as 3 hours after birth (Johnson & Morton, 1991). This talk surveys solutions attempted by computer vision researchers, placing them in the context of the biological literature. Linear principal components analysis (eigenfaces) and nonlinear artificial neural networks have been successful in pilot studies for finding and recognizing faces, lip reading, sex classification, and expression recognition. The representation of faces in these networks have global properties similar to those of neurons in regions of the primate visual cortex that respond selectively to faces. This approach depends on the availability of a large number of labelled exemplars of faces that can be used for training networks and extracting statistical properties from the sample population.

Presenters: T. J. Sejnowski (Neural Networks) and A. Pentland (Finding, Organizing, and Interpreting Faces)

Introduction

There is a long history of research into face recognition and interpretation. Much of the work in computer recognition of faces has focused on detecting individual features such as the eyes, nose, mouth, and head outline, and defining a face model by the position, size, and relationships among these features. Beginning with Bledsoe's (1966) and Kanade's (1973, 1977) early systems, a number of automated or semi-automated face recognition strategies have modeled and classified faces based on normalized distances and ratios among feature points such as eye corners, mouth corners, nose tip, and chin point (e.g. Goldstein et al., 1971; Kaya & Kobayashi, 1972; Cannon et al., 1986; Craw et al., 1987). Recently this general approach has been continued and improved by the work of Yuille and his colleagues (Yuille, 1991). Their strategy is based on "deformable templates", which are parameterized models of the face and its features in which the parameter values are determined by interactions with the image.

Such approaches have proven difficult to extend to multiple views, and have often been quite fragile, requiring a good initial guess to guide them. In contrast, humans have remarkable abilities to recognize familiar faces under a wide range of conditions, including the ravages of aging. Research in human strategies of face recognition, moreover, has shown that individual features and their immediate relationships comprise an insufficient representation to account

for the performance of adult human face identification (Carey & Diamond, 1977). Nonetheless, this approach to face recognition remains the most popular one in the computer vision literature.

In contrast, recent approaches to face identification seek to capture the configurational, or gestalt-like nature of the task. These more global methods, including many neural network systems, have proven much more successful and robust (Cottrell & Fleming, 1990; Golomb et al., 1991; Brunelli & Poggio, 1991; O'Toole et al., 1988, 1991; Turk & Pentland, 1989, 1991). For instance, the eigenface (Turk & Pentland, 1991) technique has been successfully applied to "mugshot" databases as large as 8,000 face images (3,000 people), with recognition rates that are well in excess of 90% (Pentland, 1992), and neural networks have performed as well as humans on the problem of identifying sex from faces (Golomb et al., 1991).

Important Subproblems

The problem of recognizing and interpreting faces comprises four main subproblem areas:

- Finding faces and facial features. This problem would be considered a segmentation problem in the machine vision literature, and a detection problem in the pattern recognition literature.
- Recognizing faces and facial features. This problem requires defining a similarity metric that allows comparison between examples; this is the fundamental operation in database access.
- Tracking faces and facial features. Because facial motion is very fast (with respect to either human or biological vision systems), the techniques of optimal estimation and control are required to obtain robust performance.
- Temporal interpretation. The problem of interpretation is often too difficult to solve from a single frame and requires temporal context for its solution. Similar problems of interpretation are found in speech processing, and it is likely that speech methods such as hidden Markov models, discrete Kalman filters, and dynamic time warping will prove useful in the facial domain as well.

Major Approaches

There are three basic approaches that have been taken to address these problems:

View-based approaches. This class of methods attempts to recognize features, faces, and so forth, based on their 2D appearance without attempting to recover the 3D geometry of the scene. Such methods have the advantage that they are typically fast and simple, and can be trained directly from the image data. They have the disadvantage that they can become unreliable and unwieldy when there are many different views that must be considered.

Volume-based approaches. This class of methods attempts to interpret the image in terms of the underlying 3D geometry before attempting interpretation or recognition. These techniques have the advantage that they can be

extremely accurate, but have the disadvantage that they are often slow, fragile, and usually must be trained by hand.

Dynamic approaches. These techniques derive from speech and robotics research, where it is necessary to deal with complex, rapidly evolving phenomena. As a consequence, methods such as hidden Markov models and Kalman filtering are applied in order to allow integration of sensor evidence over time, thus making possible reliable, real-time estimation.

Biological Foundation of Face Processing

Before summarizing the methods that have been devised for recognizing and interpreting faces in computer science, it is worthwhile to explore our current understanding of biological visual systems. Face recognition has important biological significance for primates, and facial expressions convey important social information. The scientific problem of how nature solves these problems is central to this report, and we may learn something from biological solutions that will help us in designing machines that attempt to solve the same problem (Churchland & Sejnowski, 1992).

Neurons in early stages of visual processing in visual cortex respond to specific visual features, such as the orientation of an edge, within a limited region of the visual field (Hubel & Wiesel, 1962). The response of a neuron in the primary visual cortex is broadly tuned to parameters such as orientation and color. The information about the shape of an object is distributed over many neurons. In intermediate stages of visual processing, neurons respond over larger regions of the visual field and they respond with greater specificity.

In monkeys, neurons have been found in temporal cortex and the amygdala that respond selectively to images of faces (Desimone, 1991). Such cells respond vigorously to particular faces but little, if at all, to other simple and complex geometric objects. No single feature of a face, such as the eyes or mouth, is necessary or sufficient to produce the response. Such neurons are good candidates for coding the properties of individual faces because their response is invariant over a wide range of spatial locations, a range of sizes of the face, and, to a more limited extent, the orientation and pose of the face. Recent statistical analysis of neurons in the temporal cortex suggest that they may be coding the physical dimensions of faces, whereas the responses of neurons in the amygdala may be better characterized by the social significance or emotional value of the face (Young & Yamane, 1992). Others areas of visual cortex are involved in representing facial expressions.

It is likely that specialized areas exist in the visual cortex of humans that are similar to those found in other primates. Humans with bilateral lesions in the mesial aspect of the occipitotemporal junction have selective deficits in recognizing faces of even familiar individuals (Damasio et al., 1982; Tranel et al., 1988). Lesions in other areas lead to deficits in recognizing facial expressions, but not recognition of identity. Although these data are indirect, they are consistent with the hypothesis that primates have special purpose cortical areas for processing faces.

Face Detection

Detection and location of faces in images encounter two major problems: scale and pose. The scale problem is usually solved by forming a multi-resolution representation of the input image and performing the same detection procedure at different resolutions (Burt, 1988b; Perry & Carney, 1990; Viennet & Fogelman-Soulie, 1992). Pose is a more difficult problem, and currently methods employing representations at several orientations are being investigated, with promising early results (Bichsel & Pentland, 1992).

Strategies in face detection vary a lot, depending on the type of input images. Posed portraits of faces with uniform background constitute the majority of current applications. In this very simple situation, face detection can be accomplished by simple methods. For instance, face edges can be found (Brunelli, 1991; Cannon et al., 1986; Wong et al., 1989), or eyes, nose, mouth, etc., can be found using deformable templates (Yuille, 1991) or sub-templates (Sakai et al., 1969). Even the simplest histogramming methods (summing the image intensity along rows or columns) have been successfully used in this simple situation (Kanade, 1973; Brunelli, 1990).

However, such simple methods have may have difficulties when facial expressions vary. For example, a winking eye or a laughing mouth can pose a serious problem. Moreover, they are simply not adequate to deal with more complex situations.

Detection of faces in images with complex backgrounds requires a strategy of a different kind. Often it is a good idea to start with simple cues such as color (Satoh et al., 1990) or motion (Turk & Pentland, 1991) to locate the potential face targets for further verification. Such initial coarse detection has the effect of greatly reducing processing expense.

These initial quick detection methods, must then be followed by more precise and reliable methods. This problem has been approached in three ways: feature-based templates, intensity-based templates, and neural networks. In the feature-based template approach, features such as the left-side, the right- side, and the hair/top contours, are extracted and grouped and matched to a template face (Govindaraju, 1992). In the intensity-based template approach, principle components of face images are used to locate the potential face regions (Turk & Pentland, 1991). In the neural network approach, face examples and background examples are used to train the neural network, and it is then used to locate candidate faces (Viennet & Fogelman-Soulie, 1992). These methods may be combined with optical preprocessing to obtain very fast face detection (see Tutorial on Hardware, pages 29-31; Wang & George, 1991).

Tracking head motion has been studied mostly under the assumption that the background is either stationary or uniform. Taking the difference between successive frames will thus locate the moving head or body (Turk & Pentland, 1991). Features on the head, such as the hair and face, are then extracted from motion-segmented image region (Mase et al., 1990). Alternatively, one can put marks, say blue dots, on several key points on the face and then the system can track the head motion by extracting the marks (Ohmura et al., 1988).

Several systems require interactive extraction of facial features on the first frame of an image sequence. The systems can then track the motion of these features using a model of the human head or body (Huang et al., 1991; Yamamoto & Koshikawa, 1991).

Human body motions are highly constrained and therefore can be modeled well by a few parameters. Systems that track body motion by constrained motion (O'Rourke & Badler, 1980), Kalman filtering (Pentland & Sclaroff, 1991; Azarbayejani et al., 1992), and Hidden Markov Model (Yamato et al., 1992), have been demonstrated. The most precise tracking reported had a standard deviation error of approximately 1 centimeter in translation and 4 degrees in rotation (Azarbayejani et al., 1992), obtained using a Kalman filter approach.

Finally, face or head tracking can be done by performing fast face detection on each frame. Multi-resolution template matching (Burt, 1988b; Bichsel & Pentland, 1992) and optical transform (see Tutorial on Hardware, pages 29-31) are three such examples. The most precise tracking reported using this approach had a standard deviation error of approximately 2 centimeters in translation (Bichsel & Pentland, 1992).

Face Recognition

One relatively successful approach to face recognition (detection of identity from a set of possibilities) is one that extracts global features from 2D, static images. The techniques use principal components analysis of the images, whether directly or via a neural network implementation (Cottrell & Fleming, 1990; Golomb et al., 1991; O'Toole et al., 1988, 1991, 1993; Turk & Pentland, 1989; 1991). For recognition, the projections of new faces onto these principal components are compared with stored projections of the training faces, and are either correlated or compared more non-linearly with a neural net. The extracted features have been called eigenfaces or holons. These "features" look like ghostly faces, and can be thought of as a weighted template matching approach using multiple templates extracted from the data.

These approaches have been tried on carefully controlled databases with about 16 to 20 subjects, yielding recognition rates of approximately 95 to 100% (Cottrell & Fleming, 1990; Turk & Pentland, 1991). More recently, they have been successfully applied to "mugshot" databases as large as 8,000 face images (3,000 people), with recognition rates that appear to be well above 90% (Pentland, 1992).

In controlled tests, these approaches have been found to be insensitive to facial expression and lighting direction (but not to shadowing) (Turk & Pentland, 1991). However, they are sensitive to orientation and scale changes, with scale being the most important, followed by orientation. Scale may be solved if one can scale the face to fit the templates in advance, or equivalently, by storing templates at multiple scales (Burt, 1988b; Bichsel & Pentland, 1992). Orientation appears to require multiple templates for this approach to work (Bichsel & Pentland, 1992). It is important to determine how the number of required templates scales as the number of subjects is increased.

On the other hand, these approaches appear quite robust to occlusions, and this simple technique may be capable of approaching human levels of performance, depending on the storage available for the various templates representing the conditions. They have been applied with limited success to expression recognition (Cottrell & Metcalfe, 1991; Turk, 1991), the templates can easily be used to detect the location of faces in the image (Turk & Pentland, 1991) (see previous section), and finally, templates of parts of the face, such as "eigeneyes", may be used to verify the match or detect important features such as gaze angle or blink rate (Turk, 1991).

The idea of global analysis using an eigenvector basis has been extended to 3D by Pentland and Sclaroff (1991). The major problem in this approach is to relate 2D and 3D information back to some canonical 3D representation. Classically, this can be solved by the technique of Galerkin projection, and is the basis of the well-known finite element method. In Pentland's method, a set of "eigenshapes," analogous to the 2D eigenfaces or holons discussed above, were created using a finite element model of compact, head-like shapes. In this approach shape is described as some base shape, e.g., a sphere, that has been deformed by linear superposition of an orthogonal set of deformations such as stretching, shearing, bending, etc. This set of orthogonal deformations are the eigenshapes, and form a canonical representation of the 3D shape.

To describe a shape, the 2D or 3D data is projected onto these eigenshapes, to determine how much of each deformation is required to describe the shape. The coefficients obtained describe the object uniquely, and may be used to compare the object's shape to that of known objects.

Experiments using this approach to recognition have involved eight to sixteen people, and have used either silhouettes or range data as input. Recognition accuracies of approximately 95% have been achieved (Pentland & Horowitz, 1991). One of the most interesting aspects of this approach is that this accuracy seems to be independent of orientation, scale, and illumination.

Deformable templates (Fischler & Elschlager, 1973; Yuille et al., 1989; Yuille, 1991; Buhmann et al., 1989) are another approach that appears very promising. Yuille constructs analytic templates of face features and parameterizes them. The parameters are then used to define a Lyapunov function which is minimized when a match is found. The system thus does gradient descent in the parameters of the templates to detect the features. By ordering the weightings of the parameters in successive minimizations, a nice sequential behavior results in which first the eye is located, then the template oriented, and finally the fine matching of features is performed. This approach is subject to local minima in the Lyapunov function, but a more sophisticated matching strategy avoids this problem (Hallinan, 1991). Robust matching methods may be used (McKendall & Mintz, 1989) to give some ability to deal with occlusions (Yuille & Hallinan, 1992). A disadvantage of this kind of method is the amount of calculation required, five minutes processing time on a SUN, to detect the features. An advantage is that careful choice of the parameters makes the approach insensitive to scale.

Buhmann and colleagues (Buhmann et al., 1989, 1991), for instance, use a

deformable template approach with global templates (see also Fischler & Elschlager, 1973). A grid is laid over an example face, and Gabor jets (a set of coefficients of Gabor filters of various orientations, resolutions and frequencies) are extracted at each grid point. So again, this is an unsupervised technique for extracting features from the data. The features are global at the level of Gabor jets because nearly the whole face can be regenerated from one of them.

Given a new face, the grid is deformed using an energy minimization approach (similar to Yuille's technique) until the best match is found. This results in the ability of the system to deal with orientation changes by producing the best match with the deformed template, and only one "training example" is necessary for each person. The disadvantage is that the system must potentially check the match to every stored template (corresponding the number of known faces) although it is likely that efficient data structures could be designed to store similar faces together. In current work, they have achieved an 88% recognition rate from a gallery of 100 faces (v.d. Malsburg, personal communication).

Burt (1988a, 1988b) uses a resolution hierarchy approach with specialized hardware to locate the face in the image at low resolution, and then proceeds with matching at higher resolutions to identify the face. At each stage, progressively more detailed templates are used in the matching process. This approach is promising because the efficient use of the pyramidal image representation and hardware allows near real-time face identification. Recent work by Bichsel and Pentland (1992) have extended this approach to include orientation (by using whole-face templates), and have been able to achieve matching rates of up to 10 frames per second on an unaided Sun 4 processor.

Tracking Faces

Face motion produces optical flow in the image. Although noisy, averaged optical flow can be reliably used to track facial motion. The optical flow approach (Mase & Pentland, 1990a, 1991; Mase, 1991) to describing face motion has the advantage of not requiring a feature detection stage of processing. Dense flow information is available throughout the entire facial area, regardless of the existence of facial features, even on the cheeks and forehead (Mase, 1991). Because optical flow is the visible result of movement and is expressed in terms of velocity, it is a direct representation of facial actions (Mase & Pentland, 1990a, 1991). Thus, optical flow analysis provides a good basis for further interpretation of facial action. Even qualitative measurement of optical flow can be useful; for instance, we can focus on the areas where nonzero flow is observed for further processing, and we can detect stopping and/or the reversal of motion of facial expressions by observing when the flow becomes zero (Mase & Pentland, 1990a, 1991).

Lip reading

Visible speech signals can supplement acoustic speech signals, especially in a noisy environment or for the hearing impaired (Sumby & Pollack, 1954). The face, and in particular the region around the lips, contains significant phonemic and articulatory information. However, the vocal tract is not visible and some

phonemes, such as [p], [b] and [m], cannot be distinguished.

Petajan (1984), in his doctoral dissertation, developed a pattern matching recognition approach using the oral cavity shadow of a single speaker. His system measured the height, area, width, and perimeter of the oral cavity. Brooke and Petajan (1986) used a radial measure of the lip's motion to distinguish between phonemes and to synthesize animation of speech. Petajan improved his system at Bell Laboratories (Petajan et al., 1988), using only the easily computable area feature and a set of acoustic rules, to achieve near-real-time performance.

Nishida (1986), of the MIT Media Laboratory, used optical information from the oral cavity to find word boundaries for an acoustic automatic speech recognizer. Nishida's work was the first one that used dynamic features of the optical signal. Nishida found that the derivative exceeded a given threshold at a word boundary, since changes in dark areas are abrupt as the pace of speech articulation is interrupted at a word boundary.

Pentland and Mase (1989) and Mase and Pentland (1991), also at MIT the Media Laboratory, were the first to use a velocity or motion-based analysis of speech. Their technique used optical flow analysis, followed by eigenvector analysis and dynamic time warping, to do automatic lipreading. They were able to achieve roughly 80% accuracy for continuously-spoken digits across four speakers, and 90% accuracy when voicing information was available. Perhaps the most interesting element of this research was the finding that the observed "eigenmotions" corresponded to elements of the FACS model of face motions. Mase (see "Tracking Faces" above and page 29) was later able to extend this approach to recognizing a wide range of facial motions and expressions.

Stephen Smith (1989) reported using optical information from the derivatives of the area and height features to distinguish among the four words that an acoustic automatic speech recognizer confused. Using the two derivatives, Smith could distinguish perfectly among the four acoustically confused words.

Garcia and Goldschen, using a synchronized optical/acoustic database developed by Petajan of 450 single-speaker TIMIT sentences, have analyzed -- by means of correlation and principal component analysis -- the features that are most important for continuous speech recognition (Garcia et al., 1992). The unit of optical speech recognition is the "viseme," a term coined by Fisher (1968), which stands for the (distinguishable) particular sequence of oral cavity region movements (shape) that corresponds to a phoneme. The novelty of their approach is that the techniques of feature extraction pointed to some unexpected grouping of correlated features, and demonstrated the need to put particular emphasis on the dynamic aspects of some features.

Psychoacoustic experiments on humans strongly suggest that the visual and acoustic speech signals are combined before the phonemic segmentation. Ben Yuhas designed a neural network to map normalized images of the mouth into acoustic spectra for nine vowels (Yuhas et al., 1989). The goal of his research was to combine the optical information with the acoustic information to improve the signal-to-noise ratio before phonemic recognition. As acoustic recognition

degraded with noise, the optical system for recognition maintained the overall performance. For small vocabularies (ten utterances) Stork, Wolff and Levine (1992) have demonstrated the robustness of a speaker-independent time-delay neural network recognition of both optical and acoustic signals over a purely acoustic recognizer.

Neural Networks

Neural networks and the eigenface approach hold promise for producing computation-efficient solutions to the problem of recognizing facial expressions. In this section we present an introduction to feedforward neural networks. Networks provide nonlinear generalizations of many useful statistical techniques such as clustering and principal components analysis. Preliminary results of the application of neural networks to sex recognition and facial expressions are encouraging. They also can be implemented on parallel computers and special purpose optical and electronic devices (see Tutorial on Hardware, page 29-31) so that relatively cost-effective solutions to the real-time analysis of faces are in the offing (Sejnowski & Churchland, 1992).

Neural networks are algorithms, inspired more or less by the types of computational structures found in the brain, enabling computers to learn from experience. Such networks comprise processing elements, known as "units", which are analogous to neurons. These are classed as input units, hidden units, and output units. One unit connected to another implies that activity of one unit directly influences the activity of the other; the propensity of activity in one unit to induce or inhibit activity in the other is called the "weight" of the connection between these units. Networks learn by modifying these connection strengths or "weights."

Input units, akin to sensory receptors in the nervous system, receive information from outside the network. In the nervous system, a sensory receptor must transduce a signal such as light intensity or pressure into the strength of a signal; in neural networks the strength of the input signals is determined by the nature of the problem. In the case of vision, the inputs might be a gray-level array corresponding to the image being classified, or a processing version of the inputs, which might include feature extractions. (If the feature extraction is a nonlinear process then important information may be removed from the input, and the performance may be degraded. For an example of this see the section below on sex recognition.) The input signal is relayed from the input unit to a hidden unit. (Input units may alternatively send their signals directly to output units, but the class of problems which can be solved using this technique is more limited; it corresponds to problems termed "linearly separable", meaning that if the input points were graphed with relevant axes, it would be possible to draw a line separating the points in each output class from those in the others -- or if the points are in a space of n dimensions, a hyperplane of n-1 dimensions could be made to separate the output classes. Input units can, however, be made to send their signals both to hidden and output units leading to more complex network architectures.) Hidden units, analogous to interneurons, serve solely as intermediate processors; they receive signals from input units and send signals either to further layers of hidden units or to output units. Their job is to perform a nonlinear transformation of the inputs, making a three-layer

network more powerful than a two layer network without hidden units. Output units serve up the outcome of the processing. Thus, they can express anything from how strongly a muscle fiber should twitch (motor neurons in biology are classic output units) to the recognition of the expression on a face.

The "architecture" of a network comprises the details of how many layers are present, how many units invest each layer, and which units are connected to which others. In a standard three-layer network, there is one input layer, one hidden layer, and one output layer. In a fully connected feedforward network, all input units connect to all hidden units, and all hidden units connect to all output units; however many variations on this theme exist. This class of networks is called feedforward because activity in a unit only influences the activity of units in later layers, not earlier ones; feedback or recurrent networks can also be constructed and are discussed in the next subsection. For each connection between two units, a "weight", akin to a synaptic efficacy, characterizes the "strength" of a connection -- the propensity of one neuron to cause the neuron to which it feeds to become active. Learning by the network requires the selective modification of these weights, and different strategies have been devised to accomplish this. Again, networks "learn" by successively modifying the strengths of the connections between units, in a direction to reduce the error at the output.

Backpropagation of Errors

Backpropagation represents one much used strategy for computing the gradients of the weights with respect to the overall error (Rumelhart et al., 1986). The weights, initially set to small random values, are iteratively changed to reduce the error of the output units for each input pattern. For a given input pattern, the activity of each hidden unit, and later of each output unit, is calculated. The output function (chosen by the network architect) determines the activity of a unit based on this weighted summed input, and is usually taken as a nonlinear sigmoid function. With this choice the output cannot increase without bound as the incoming signals and weights increase. In statistics, fitting input data with sigmoid functions leads to nonlinear logistic regression. Many other nonlinear output functions can also be used and the choice is dictated by practical issues such as speed of training, accuracy, and amount of available data.

For each training example (which serves as an "input" to the network, i.e. a full set of activities for the input units), the actual network output (the collective activities or outputs of all "output" units) is compared to the desired output and an error is calculated. A summed (squared) error across all training examples is obtained, and by taking the derivative of this error with respect to a given weight, one can determine the direction to modify the weight in order to minimize the output error. The weights for hidden-to-output units are modified by a small amount in the direction to reduce the output error, and the "chain rule" from elementary calculus is invoked to extend or "back propagate" this differentiation in order to modify by a small amount the weights at the earlier input-to-hidden level. (The amount by which weights are modified is given by the "learning rate", a variable parameter.) The whole process is repeated, again giving each training example as input, calculating the output error, and

incrementally and iteratively modifying the weights until the error begins to asymptote (or until "cross validation" techniques, which involve testing the network with untrained examples, suggest that further training will yield "learning" which is not generalizable to examples outside the training set).

To make this process concrete and germane, a simple compression-net can be considered (see below). A set of face images serves as input; normalized gray level values for each point on a 30x30 pixel image provide values for each of 900 input units (each receiving information from one of the 900 points on the image, in analogy to photoreceptors in the retina). These activities are passed through initially random weights to 40 hidden units, which, in turn, are connected by initially random weights to a single output unit, which is meant to ultimately give a value of zero if the input face is female and 1 if male. The actual output will have no semblance to the desired output with the initially random weights, and the output error will be calculated. A second training face will be shown, and its output error calculated. After all the training faces have been presented (reserving some faces which the network has not trained on for testing the network later), a summed error across all faces will be calculated and the weights will be slightly modified to make the network do less badly with the next round of presentations. This process will be repeated until the network is doing as well as it seems likely to, at which point "test" faces, which the network has never trained on, can be presented to evaluate the network's performance on the task. The 40 hidden units are low-dimensional representations of the face. The weights to the hidden units look like ghosts when viewed as images.

Many of the mysteries regarding the mathematical properties of feedforward networks and backpropagation have yielded to analysis over the last few years. We now know that they are universal approximators in the sense that feedforward networks can approximate well-behaved functions to arbitrary accuracy. The complexity of feedforward network models is also well understood in the sense that the bounds on the number of training examples needed to constrain the weights in the network have been established. Networks with linear hidden units perform principal components analysis, and nonlinear hidden units provide a nonlinear generalization of this technique. These simple networks have proved their worth in diverse tasks ranging from determining whether patients presenting to an emergency room with chest pain are having a heart attack, to making currency trading decisions, to deciding whether cells under a microscope are likely to be cancerous. The key to success in all of these examples is an adequate database of training data.

The feedforward architecture used in many neural network applications is quite versatile. A variety of functions have been used for the nodes in the hidden layer in place of the sigmoid function. For example, radial basis functions have been used with effectiveness and have several advantages for some problems, including faster training (Poggio, 1990). For many problems such as speech recognition, where the information is spread out over time, the temporal patterns can be mapped into a spatial array, converting the temporal pattern into a spatial pattern -- an arrangement called a time-delay neural network (Sejnowski & Rosenberg, 1987; Waibel et al., 1989). More advanced neural network architectures incorporate dynamical properties, such as temporal

processing in the nodes (time constants or short-term memory) and feedback connections (Pearlmutter, 1989). These neural network architectures have also been used for solving control problems, such as controlling the limb of a robot arm (Jordan, 1992) or eye tracking of moving objects (Lisberger & Sejnowski, 1992).

Sex Recognition

Humans are competent at recognizing the sex of an individual from his or her face, though in real life the task may be facilitated by non-featural cues of facial hair or male-pattern baldness, by social cues of hairstyle, makeup, and jewelry, and by non-facial biological and social cues of size, body morphology, voice, dress style, and mannerisms. The task of distinguishing sex from static facial cues alone, in the absence of hairstyle, makeup and other disambiguating cues, is more difficult, though humans still perform quite well. Performance comparable to humans has been reported using neural networks (Cottrell & Metcalfe, 1991; Golomb et al., 1991; Brunelli & Poggio, 1991). These networks rely on preprocessing the image by normalizing it (for instance re-sizing and centering them), and either extracting features or "compressing" the image using autoencoding (as described above). In autoencoding, a method that is equivalent to "eigenfaces" when the hidden units are linear, a network is asked to reproduce each input image as output after forcing it through a "bottleneck" of hidden units. The hidden layer "recodes" each image with many fewer units. The new more parsimonious representation of each face, given by the activities of the hidden units of the "autoencoder" for that face, can substitute for the face -- for instance as input to subsequent networks.

Brunelli and Poggio (1991), using a radial basis functions in the hidden layer, found that the position of the eyebrow was more important than the position of the mouth for gender classification for their data set; however, the practice of brow tweezing among women suggests their network may have tapped into an artificial sign (akin to makeup and hairstyle) rather than a static or slow sign of identity (see page 9). However, comparable performance could be achieved using the gray-level representation of the face directly (Golomb et al., 1991). Recently, it has been shown that compression is not needed and good performance is possible from a two-layer network with direct connections from a normalized gray level image to a single output unit (Gray et al., 1993). This surprising result indicates that extraction of sex information in faces is less complex than had been assumed. Examination of the weights in this model reveal that information about the sex of a person is distributed over many regions of the face.

Expression Recognition

Our present experience on network architectures for expression recognition is limited. Most work has involved frontal images of static faces under good illumination (Cottrell & Metcalfe, 1991). In one unpublished, preliminary study using expressions from a single person, Golomb trained a network to recognize eight distinct facial actions. The network was trained on 9 examples (of variable intensities) of each facial action, and was tested on a tenth, different, example of that facial action; this process was iterated (using a technique termed

"jackknifing"), reserving a different face for testing each time and training from scratch on the remaining nine. The data from the ten independent networks was compiled for statistical analysis. The facial actions employed corresponded roughly, in lay terms, to smile, frown, brow-raise, sneer, squint, pucker-lips, purse-lips, and neutral expression. As expected, the two most similar facial expressions (purse-lips and pucker-lips), which were difficult for human observers to distinguish in some instances, took longer for the network to learn than more dissimilar expressions. These similar expressions were selected for the purpose of assessing how well the network would do when challenged with subtle distinctions in facial expression. However, the "neutral" expression, though never misclassified as any other expression, took longest for the network to learn, though ultimately test cases of all eight expressions were correctly categorized by the network in almost all instances. (Interestingly, human observers also had difficulty classifying neutral faces as such, though they would not classify them among any of the other available options)

Optical flow can also used to extract dynamical muscle actions from sequences of images. Facial actions forming a fifteen-dimensional feature vector was used to categorize four expressions using a nearest-neighbor technique (Mase & Pentland, 1991). The eigenface approach has also been used to successfully classify expressions for a single person (Turk, 1991). More recently, a variation in the eigenface approach has successfully classified the six basic emotional states across a database of eleven people (Pentland et al., 1992).

These preliminary results are encouraging and provide evidence that the recognition part of the problem of automating the classification of expressions may be solvable with existing methods.

NOTE:. This section was based on tutorials given by Alexander Pentland and Terrence Sejnowski at the Workshop. Beatrice Golomb assisted with writing the section on neural networks.

III-D. Special Hardware For Face Processing

Demetri Psaltis

Abstract. The optical and electronic methods for performing image processing tasks, such as face recognition, and the requirements imposed on the hardware for real time operation are described. Real time image processing is difficult because of the very high computational rate that is required to make interesting calculations on large images. Parallel hardware is an obvious solution to the problem particularly since image processing lends itself naturally to fine grain parallelism, with each pixel being processed in parallel at each stage. For certain regular and/or local operations, digital techniques can be effectively applied to image processing (e.g., cellular and systolic arrays). Problems with power consumption, topology and interconnections, however, can make analog implementations using VLSI or optical techniques advantageous. The primary focus is on such analog hardware for early vision tasks.

Presenter: D. Psaltis

Face recognition algorithms can be classified into two broad categories: model-based algorithms and learning algorithms. Model based algorithms are in general computationally intensive, requiring complex logic operations that typically require the flexibility of a general purpose digital computer. On the other hand, algorithms that are based on learning approaches, (neural networks, eigenfaces, statistical approaches, etc.), typically require relatively simple elementary operations and are conducive to massively parallel implementations. As a general rule then, approaches such as neural networks have the advantage that the speed, efficiency, and cost of the implementation can be greatly reduced by special purpose hardware. The hardware options for the implementation of a neural network face recognition system range from workstations, to general purpose supercomputers, to custom digital and analog VLSI, and optics. Workstations and, in some cases, supercomputers are used for algorithm development. Special purpose co-processors and digital signal processing cards can speed up the execution sufficiently to allow real time operation for simple recognition tasks. An excellent example in this category is Burt's (1988b) work in which a simple workstation with a special purpose digital co-processor was able to track the face of a person staring forward anywhere in the field of view of the television camera. In the work on eigenfaces by Turk and Pentland (1991), the data reduction achieved by the algorithm, makes it possible to use commercial hardware to obtain near real time performance. The implementation of more complex face processing tasks where we can have insensitivity to the illumination direction, head position, orientation and size as well as good discrimination capability will require, in most cases, special purpose hardware for real time, compact operation.

The reason for switching to special purpose hardware is not only so that we can implement more efficiently an algorithm that was previously working in software or on a slower machine. Rather, we need to select an implementation strategy that is inherently well suited to the task at hand and dramatically enhances the processing power that is available. There are two major categories of special purpose hardware that appear best suited for real time implementation of neural, image processing algorithms: analog VLSI and optics. Both of these technologies are analog. In switching from digital to analog we give up a lot. We give up algorithmic flexibility and accuracy. With neural networks, however, adaptation helps us address these problems. A neural network is programmed to perform specific tasks largely through the learning process instead of relying solely on the circuit design. Therefore, the fact that we have less control over the initial functionality of the machine is compensated by the ability to train it. Adaptation (along with redundancy) also helps overcome the accuracy limitations since an error in computation is detected and compensated for by the adaptation mechanisms that are built in. The benefits of the analog implementation are extremely dense, power efficient, and powerful networks. An analog multiplier is constructed with 3 to 5 transistors whereas an entire chip is needed for a digital multiplier. In what follows we will briefly discuss each of the two hardware technologies, analog VLSI and optics.

Analog VLSI

One of the most dramatic demonstrations of the power of analog VLSI (Mead, 1989) has been in the area of image processing. Most analog VLSI image processing chips consist of a 2D array of photosensors coupled to surrounding circuitry that performs a local computation on the image field. The first circuit of this type is the silicon retina of Mahowald and Mead (1991). In this circuit, a 2D hexagonal array of phototransistors senses the incident image. Each photodetector is coupled to its 6 neighbors through a resistive array. The circuit at each node computes a local edge enhancement and also emphasizes the time varying portion of the signal using differentiation of the signals. Several other circuits of this type have been constructed (Andreou et al., 1991; Harris et al., 1990; Tanner & Mead, 1984; Tolbruck, 1992), most significantly chips that perform motion processing. These analog VLSI chips very effectively perform preprocessing tasks that can be implemented with local connectivity. For computations that require pixels in the retina to communicate with many of their neighbors, the achievable density of image pixels deteriorates roughly as the square of the number of connections per pixel. This is because if the number of connections is doubled the area that needs to be devoted to them quadruples since we not only have more connections but also longer connections. Recently, several researchers have demonstrated analog VLSI image processing chips that have an optical output for each image pixel (Drabik & Handschy, 1990; Cotter et al., 1990). This is accomplished by depositing liquid crystal light modulators on top of the silicon chip. This provides more extensive connectivity and the capability to cascade in parallel such chips using the optical techniques we discuss below.

Optics

The advantages of the optical implementation derive from the fact that we have

direct access optically to the third dimension (Psaltis et al., 1990). This is particularly useful for image processing in general and face processing in particular, since it allows us to arrange the image pixels in 2D arrays that densely populate the plane, while making the interconnections via the third dimension. Typically, the interconnections are specified with holograms that are dynamically adapted. This basic arrangement makes it possible to process images with roughly one million pixels within an active area of approximately 1 squared cm with essentially any desired connectivity. The optical implementation becomes increasingly attractive as the number of connections per pixel increases. For applications such as motion detection and edge enhancement, which can be realized with very few local connections, a purely electronic approach may be sufficient. However, as the density and range of the connections increases, the connections start dominating the area of the chip, and the density of pixels that can be supported reduces dramatically. The pixel density that is achievable with an optical implementation is relatively insensitive to the type of connectivity that is required. A large number of optical implementations have been described and experimentally demonstrated (Psaltis & Farhat, 1985; Abu-Mostafa & Psaltis, 1987; Owechko et al., 1987; Anderson, 1986; Farhat et al., 1985; Wagner & Psaltis, 1987; Psaltis et al., 1988; Yeh et al., 1988; Paek & Jung, 1991; Maniloff & Johnson, 1990). Recently, an optical experiment was carried out specifically for face recognition (Li et al., no date). This two layer network was trained to recognize in real time (at 30 frames per second) faces under a broad range of viewing conditions while exhibiting excellent discrimination capability against unfamiliar faces. The network has up to 200 hidden units and more than 10 million adaptable weights. This is probably the most ambitious face processing hardware demonstration to date. The next step is likely to be a marriage of the silicon retinas that perform the early preprocessing followed by an optical system that performs the recognition task.

IV. REPORTS FROM THE PLANNING BREAKOUT GROUPS

IV-A. Breakout Group on Basic Science

Participants: J. Allman, J. T. Cacioppo, R. J. Davidson, P. Ekman, W. V. Friesen, C. E. Izard, M. Phillips

Basic Science for Understanding Facial Expression

Richard Davidson, John Allman, John Cacioppo, Paul Ekman, Wallace Friesen, Joseph C. Hager, and Mike Phillips

This group was given the assignment of specifying basic research needs in the emerging area of facial expression understanding. This report first presents a series of recommendations that specify the most critical basic research needs, the infrastructure that needs to be developed, and the data bases that need to be developed. This is followed by a discussion of next steps that can be taken now to reduce the time spent in facial measurement before a fully automated system for measuring facial movement is available.

Recommendations on Basic Research Needs

Perception of facial expression

Research should determine:

- whether high levels of agreement among observers can be obtained about which emotion is displayed and/or the intensity of the emotion which is displayed, without presenting the full facial configurations (prototypes) which have been studied to date;
- how blends of emotion are manifest;
- the variables that influence an observers interpretation of a facial expression;
- the mechanisms that underlie the perception of facial expressions of emotion;
- the relationship between information from facial actions and other behaviors.

Darwin's (1872) pioneering studies began a century-long debate about whether observers can accurately judge the emotion shown in a facial expression. This issue is related to the question of whether specific expressions actually do

correspond to particular emotions. Over the decades, clearer conceptualization of these two problems stripped away confounding variables such as characteristics of the elicitors, blending of two or more expressions, additions of irrelevant actions of the face or body, poor photographic and procedural techniques, and language problems to reveal agreement among observers about the emotional meanings of a small number of expressions postulated to be prototypical for each of a number of emotions (Ekman, 1972; Izard, 1971). Today, studies by several researchers (reviewed by Ekman, 1989) show that such prototypes, including happy, sad, fear, anger, surprise, and disgust, are accurately judged across many cultures.

Some research studies have examined variables that affect the interpretation of facial expression. One line of studies indicates that females tend to be more accurate in judging expressions than males, but the difference is quite small (Hall, 1978). Some work has examined the influence of the context, including the eliciting circumstance and previously seen expressions, on judgments but there is disagreement about this issue and how to properly study it (Russell, 1991a, 1991b, 1991c; Russell & Fehr, 1987; Ekman & O'Sullivan, 1988; Ekman et al., 1991a, 1991b).

Only a few provocative research findings have shed any light on the specific mechanisms for perceiving facial expression. One line of investigation has focussed on hemispheric differences in processing facial stimuli, with the argument centering on whether the right hemisphere is dominant for this task (Borod et al., 1990). Another approach is identifying specific brain centers for processing facial information, such as identity (Heywood & Cowey, 1992). Two excellent studies have recently appeared that have used positron-emission tomography (PET) to reveal the neural systems involved in discriminating the gender and identity of faces in the temporal cortex (Haxby et al., 1991; Sergent et al., 1992).

Thus far, no functional anatomical studies have been published that reveal the neural systems involved in the production or interpretation of facial expression. Such studies should be strongly encouraged. The recent development of functional MRI (magnetic resonance imaging) has enormous potential for these studies because functional MRI has higher spatial and temporal resolution than PET and because multiple studies can be conducted in the same individual since subjects are not exposed to ionizing radiation.

Twenty five years of cross-cultural research finds consistent evidence for the universal recognition of six emotions -- anger, fear, disgust, sadness, happiness and surprise. Recent evidence indicates possible additions to this list, including contempt and shame. A key issue for research is to elaborate further the basic categories of emotion recognized from the same expressions across cultures and to study any differences for specific cultures. Another issue is what variations (e.g., changes in intensity or additions of facial actions) on the full face prototype expressions for an emotion are still perceived as belonging to the same basic emotion. A related question is how expressions that contain only portions of the full prototype are judged.

Another issue barely explored is how blends of different emotions in the same

expression are perceived and judged. The effect of asymmetry of expressions on the perception of emotion expression needs further clarification (Hager & Ekman, 1985). How the perception of emotion expressions affects the perception of the expresser's other personal characteristics, such as personality, intelligence, and health, should be explored.

The temporal dynamics of expression should be examined to determine if they provide information independent of what is provided by the configurational aspects of an expression. Research should examine the relationship between the dynamics of the facial expression and the information the user is attempting to convey in spoken discourse.

Differences between voluntary and involuntary facial expressions of emotion

The facial behaviors that distinguish between false versus genuine, and more broadly between voluntarily produced and involuntary facial expressions have been explored only for happiness (Ekman et al., 1988, 1990) and need to be examined for other emotions.

We need to know if the markers identified in the extant research generalize to all expressive behavior (e.g., asymmetry of facial movement).

The relationship between facial and other signs of emotion

Much more work needs to be done to determine how emotion signs interact with other facial signs, such as signs of age, sex, race, ethnicity, and linguistically-related signals. Another related area of research is the relative contribution of face compared to other signal systems, such as body movement, voice and language, and how these systems may be integrated by the expresser and decoded by the observer.

The physiological consequences of voluntary production of facial expressions

Work began in the last decade on the consequences of voluntary production of facial expressions on behavior and physiology. A growing body of evidence indicates that such voluntary production generates subjective changes in emotional feelings, shifts in autonomic nervous system activity and alterations in central nervous system patterning (Ekman et al., 1983; Ekman & Davidson, 1992; Levenson et al., 1990). We do not know which elements of particular expressions are necessary and/or sufficient for the production of these effects, nor do we know the neural circuits which subserve these effects. Studies that combine modern neuroimaging methods with behavioral procedures are needed to address this question.

Studies of spontaneous expressive behavior in response to standardized elicitors

We have few data on the temporal dynamics of spontaneous expressive displays. Such data will be critical for both machine understanding and machine production (animation) of facial expressions. We also know little about the

range and variability of the configurations (specific muscle actions) that occur in spontaneous facial expressive behavior. How frequent are the facial prototypes of emotion in different real-world situations? It is likely that there is considerable variability in spontaneous facial behavior, across cultures and social settings, in addition to some uniformities. By studying large sample sizes under varying incentive conditions, we can begin to specify just what remains invariant within an emotion, or group of related emotions, e.g., what has been called an emotion family (Ekman, 1992a). At present, this proposal assumes an invariant core of facial actions that is preserved across different instantiations of an emotion, but we have very little data on the expressive signs of such invariance.

We also know little about the contributions of non-facial expressive behavior such as head and lip movements to the understanding of facial expressions. While emotional expressive behavior appears to be particularly dense during spoken conversation, the relation between emotional and nonemotional facial behavior has not been systematically studied.

Another issue of central importance is to determine what information is carried in physiological indices that is not available in expressive measures. If we are to use automated measures of expressive behavior to make inferences about emotional state, we must know the degree to which other nonverbal measures contribute unique variance unavailable in measures of expressive behavior. Studies that systematically examine the contributions of facial and physiological (both central and peripheral) measures to both self-report and other behavioral (or task-related) manifestations of emotion are needed.

Infrastructure Recommendations

Training

A new generation of investigators must be trained who are knowledgeable about neuroanatomy, the psychology of facial expression, computer science, and neural networks. This should be accomplished in a number of ways:

- Post-doctoral research fellowships providing interdisciplinary training that uses the resources of multiple institutions.
- Summer Institutes, bringing together faculty from diverse disciplines.
- Centers of Excellence, which represent geographic centers where high concentrations of relevant investigators are present and that could be combined with post-doctoral training.
- Special journal sections to bring information from different disciplines to the attention of other relevant disciplines (e.g., have computer vision and neural network experts write a series of papers for a psychology journal and vice versa). The purpose would be to catalyze a dialogue.

Instrumentation

Appropriate video recording technology to make a psychologist's recording usable to the computer vision community was emphasized.

Standardization of inexpensive hardware platforms for digitally processing facial

images and analyzing them.

Mechanisms to facilitate the sharing of software across laboratories and training personnel in its use.

Database recommendations

The following are important to include in a database for the face:

- Compilation of extant findings with more fine-grained description of facial behavior. Virtually all of the extant findings on facial expression and emotion have not included the actual FACS codes that were used to identify particular expressions. The compilation of this information would provide some of the information that is currently lacking on the range and variability of spontaneous facial behavior. Very large data sets on adult facial behavior exist at the University of Washington (J. Gottman), University of California, San Francisco (P. Ekman), University of Saarlandes (R. Krause), Wurzberg University (H. Ellgring), University of Zurich (E. Banninger-Huber), and the Ludwig-Boltzmann Institute of Humanethologie (K. Grammer).
- Dynamic voluntary productions of facial actions, with descriptive tags, accompanied by data from other sensors (e.g., EMG, facial thermography, central nervous system measures).
- Spontaneous behavior with tags, along with associated physiology. Expressive behavior to be obtained under several conditions including interactions with both people and machines. The latter is required since it is not clear if people will interact with even intelligent machines in the same way that they interact with humans. Ideally, these interactions would include both acoustic and visual data. To be useful for studying how people would provide visual information cues when speaking to a machine, data should be collected with either a simulation of an intended spoken language system or with some initial version of such a system. Since we expect the benefits of visual cues to become more pronounced in noisy environments, at least some of these data should be collected with varying amounts of environmental noise. Not only will this alter the acoustic data, but it is likely that it will alter the behavior of the subject. Data collected in this way, would be useful for answering some basic questions about how people will interact with intelligent machines in different situations, as well as provide a means of training and testing systems to make use of the visual cues that people provide.
- Animation exemplars of various combinations of facial actions with variations in time course, for use in perception studies.

Next Steps towards Automating Facial Measurement

While all agreed that automating the entire process of facial coding would be enormously beneficial, we also recognized the likelihood that such a goal was in the relatively distant future. Accordingly, we discussed the advantages of various interim solutions that represent efforts to partially automate the system. Below we consider the relative advantages and disadvantages of different partial automated systems.

A fully automatic system may be too difficult to ever develop because of the many potential artifacts that could interfere with measurement. However it is within the range of current technology for much of the tedious and time-consuming parts of FACS scoring to be automated, freeing valuable time of trained human observers to make the most difficult judgments. Since the discovery of new phenomena may not be automatable, it is essential for humans to remain "in the loop" in any case.

Detecting when a person is speaking

Facial behavior usually occurs most often during speech. Most of these facial actions are conversational signals (e.g. movements which punctuate speech; see Ekman, [1979] for a listing of various conversational signals), rather than signs of emotions. The emotion relevant facial movements may also be most frequent when a person is speaking. A system that could identify when a person is speaking would flag locations during a conversation where there is a high likelihood of frequent facial behavior, and that could then be scored by a human coder. Speaking could be detected from a voicing detector or by identifying lip movements associated with speech. It will be important to distinguish those lip movements required for speech articulation from additional movements of the lips which are signs of emotion. This is relatively easy for a human observer to distinguish. However, the speech required lip movements do vary with particular languages. Every system described below should have this capacity as it is likely that many investigators will want to examine the relationship between speech interaction patterns and facial behavior.

Detecting head and/or eye movement or position change

Data could be provided for studies that require knowing if a person was facing and/or looking at a particular person, object or area in the visual space. Noise in such measurement would come from quick movements to and from the person's usual position, such actions as nodding 'yes' or shaking the head to indicate 'no'. Perhaps these could be identified separately by the fact that the direction of movement changes rapidly.

A system could be developed which detected when the head was turned so far away that facial muscle activity could not be scored. Time would be saved by removing those periods from the corpus which the human must then score.

In addition to methodological benefits of this type of automated system, such a system could provide data relevant to attention, and perhaps relevant to satisfaction or interest.

Detecting some of the most frequent facial actions

The evidence to date suggests that when people are engaged in a conversation, the most frequently occurring movements are brow raise, brow lower, and some form of smiling.

Brow Movement Detector

The simplest system would detect just brow raising and lowering. Some investigators might find this detector useful as a means of identifying and setting aside what they do not want to have human scorers analyze. Others might find brow raising and lowering of substantive interest, as there is some evidence that changes in the frequency of these actions are relevant to involvement in the speech process, and may also provide information about other cognitive states (see Ekman, 1979).

Measurement of brow raising and lowering should be combined with the detection of speaking and face direction described above, as there is some evidence to suggest that the signal value of these brow actions varies with whether the person is speaking or not, and with whether they are facing the other interactant or not (Chesney et al., 1990).

A more elaborate system would identify the other five movements of the eyebrows that are possible. These occur at a much lower frequency, but three of the movements are relevant to identifying the occurrence of fear or sadness.

Smile Detector

There are numerous studies where the automatic detection of the frequency and duration of the contraction of the zygomatic major muscle would provide complete or at least sufficient data. Overall smiling rate could be used to indicate satisfaction with certain types of interaction or pleasure with the person in the interaction. It will be important to distinguish the zygomatic major smile from the risorious muscle smile, as the latter has been most often found as a sign of fear.

There are now more than a dozen studies (reviewed in Ekman, 1992b) which show that zygomatic major muscle smiling is not a sign of actual enjoyment unless part of the muscle which orbits the eye (orbicularis oculi, pars medialis) is also active. It will be much more difficult to detect the presence of this action in addition to zygomatic major. It may be sufficient to simply identify that zygomatic major smiling has occurred, and then have a human scorer make the decision about the presence of the orbicularis oculi muscle.

Detecting facial movement which is neither brow action nor smiling.

If smiling and brow movement could be automatically detected, it would be very useful if a system could also detect any other facial movement apart from those actions, even if it could not discriminate among those movements. A human scorer would then inspect and score those movements. Although it may be obvious, such a system would enormously reduce the time now consumed in finding and scoring less frequent, but important, facial actions.

Detecting the onset, apex, and offset of any given facial movement.

The most time costly aspect of current facial scoring is to obtain these time markers. This information is crucial for coordinating facial activity with simultaneous changes in physiology, voice, or speech. It is also thought likely that information about the time course of a facial action may have psychological

meaning relevant to the intensity, genuineness, and other aspects of the expresser's state. And, time course information is necessary to provide a database for those wanting life-like facial animation.

The simplest system would take input from a human scorer about the particular action which had been identified, and then automatically identify the start, apex, and end points of that action. More sophisticated systems would measure different aspects of the timing of the onset to apex, and the offset.

Detecting a limited number of specified facial actions.

There are many studies in which the investigator can specify *a priori* the particular actions of interest. A system could be developed which was capable of detecting just instances of actions for which a number of exemplars were provided. For example, in studies of depression, it may be sufficient to detect any instance of sadness.

Automating Other Relevant Data Sources:

Automated gesture recognition

Although considerably less work has been performed on gesture than on facial expression, a system that automatically recognized gesture, or a partial system similar to those described for the face above, would help considerably in using both facial and gestural information together to make predictions about emotion and other behavior (see Rosenfeld, 1982, for a review on systems for measuring bodily movement and posture).

Physiological pattern recognition

The use of multiple measures of peripheral and central psychophysiological measures necessitates ways of meaningfully integrating across many measures to describe coherent patterns. Most of the extant research on multiple physiological indicators of emotion has used relatively crude procedures for characterizing patterns of physiological response. We need better statistical tools for pattern description and tests to differentiate one pattern from another (e.g., how do we know when we have a different pattern versus a variant of the same pattern?).

NOTE: Prepared by Richard J. Davidson with contributions from John Allman, John Cacioppo, Wallace Friesen, Joseph Hager and Mike Phillips. This was then substantially revised and edited by Paul Ekman.

IV-B. Breakout Group on Sensing and Processing

Participants: A. Yuille, A. Pentland, T. S. Huang, P. Burt, G. Cottrell, O. Garcia, H. Lee, K. Mase, T. Vetter, Z. Zhang

Sensing and Processing

Alan Yuille, Alexander Pentland, Peter Burt, Gary Cottrell, Oscar Garcia, Hsien-che Lee, Kenji Mase, and Thomas Vetter

Introduction and Overview

The charter of Group 2 was to investigate sensing and processing techniques for automatically extracting representations of faces and facial features from real images. These representations should be transformed into descriptions that the psychologists in Group 1 (Basic Science) consider necessary for understanding facial expressions. The goal would be to develop a fully automated system.

There has been very little work done on this problem. This report attempts to summarize the state of the art and to describe promising directions to pursue. Many of the techniques described were developed for the related, and far more studied, problem of face recognition. There now exist face recognition techniques that will reliably recognize faces under restricted viewing conditions.

Lip reading is also a closely related problem. It involves extracting and describing the motion of the lips during speech, which can be considered facial feature understanding.

We organize this section of the report as follows. First, we describe the issues of sensing and environment. Next, we investigate methods for detecting the presence of faces in the image. After that, we consider how to detect facial features first from static images and then from motion sequences assuming that the head has been located. Finally, we consider how to interpret expressions.

There are some reliable techniques for locating faces in images. There is, however, only preliminary work on detecting facial features -- though several directions seem promising. Interpreting expressions is extremely preliminary. The difficulty of all these problems depends strongly on the viewing conditions, the orientation of the head and the speed of head movement. Controlling these factors, when possible, should considerably simplify the problems.

One major difficulty in building a fully automatic system is that many of the basic science questions have not yet been addressed, such as what the

relevant patterns of facial movement are, what the combinations of these movements signify, and so forth. Input from the psychologists in Group 1 (Basic Science) and the computer animators in Group 3 (Modeling and Database) would be very useful for determining the quantitative descriptors of facial expressions, including their time dependent behavior.

There are many potential applications for a fully automatic facial feature understanding system, (e.g., performance monitoring, communications, teleconferencing, lip reading, medical diagnosis, security/intelligence, content based image processing, human-computer interactions, virtual reality, expression tracking, lip reading, animation, multi-media). The use of passive vision is desirable since it is non-invasive and could work in a large variety of environments. In certain situations it might be supplemented by active processes (see pages 39-40).

Of particular interest are human/machine interfaces. The availability of reliable head tracking, face recognition, and expression recognition systems would allow major improvements in human/computer interactive systems. The issue of what intermediate systems are worthwhile is discussed in the Group on Basic Science report.

Sensing and Environments.

Various factors must be considered when selecting and arranging the sensors for monitoring facial expression. The essential parameters are quite simple: the spatial and temporal resolution of the video images obtained, and the camera's field of view. The sensor must provide sufficient detail to discriminate expressions of interest, and it must provide a sufficiently wide field of view to ensure that the face stays in view. In general, however, meeting these basic requirements with a single fixed camera can be difficult. Research challenges in the area of sensing and environments relate to strategies for controlling camera gaze and zoom, in order to effectively extend the field of regard while maintaining high resolution.

While it may be desirable to use cameras with both as high a resolution and as wide a field of view as possible, this can place an undo burden on the computer that must analyze the resulting data. The sensor data rate is the product of field of view, spatial resolution (samples per unit angle), and temporal resolution (frame rate). Required rates depend on application, but can easily exceed practical limits on computing devices. Strategies that allow a small field of view camera to survey a wide field of regard also make most effective use of limited computing resources.

Current status

A standard NTSC video camera provides an image that, when digitized, measures 768 by 480 pixels. For a typical face monitoring task it may be necessary to arrange the camera so that there are at least 50 pixels across the width of a subject's face. The field of view can then be about ten times the width of the face. This camera should be sufficient for applications in which the subject is seated but otherwise is free to move his head. On the other hand it

may not be sufficient for applications in which the subject is free to walk in front of and approach or move away from the camera. (Note that behavioral scientists often try to fill the frame with the face to make FACS scoring easier.)

The temporal frame rate required for monitoring facial expressions depends on the types of expressions that are of interest. Some expressions, such as a smile or frown, may persist for several seconds. A frame rate as low as one frame per second may suffice if one needs only to determine presence as opposed to temporal information. Monitoring more subtle or fleeting expressions may require ten or more frames per second. Lip reading almost certainly requires full NTSC frame rates (30 frames or 60 fields per second).

Large format cameras are available. Kodak, for example, markets a camera that measures 1320 by 1035 pixels, and another that measures 2048 by 2048. However these cameras provide only 10 frames and 5 frames per second, respectively. High definition cameras are roughly 1440 by 860 pixels, and operate at full frame rates, but these are very expensive. A less expensive alternative, if a wide field of view is required, is simply to use several NTSC cameras each covering a portion of the scene.

The task of automatically tracking faces or facial features can be simplified considerably through the addition of marks on the face. While it is desirable to monitor faces as unobtrusively as possible, the use of facial markings may be expedient in the near term for research applications, such as the study of facial expression and the development of computer interfaces that monitor the user's face.

TV cameras can be augmented with other sensors to obtain additional information about a face. For example, sensors have been developed that provide 3D range data. Commercially available range systems are too slow to be used in monitoring expressions. But new devices are being built that have the potential of providing an updated range map at frame rate, 30 frames per second.

Key research challenges

As noted above, a single camera can pan and zoom under computer control to follow faces as they move. This strategy can, in effect, provide a very wide field of view at high resolution, while keeping data rates and computation loads low. But use of a controlled camera introduces other complications. A special camera mount with drive motors is required. And fast image analysis is required to determine where to orient the camera on a moment by moment basis. The development of sensors and analysis techniques with these capabilities is the subject of research in the field of "active vision".

In general terms the objective of active camera control is to focus sensing resources on relatively small regions of the scene that contain critical information. However, a vision system often must also observe the scene with a wide field of view camera (at the low resolution) in order to determine where to direct the high resolution observations. This is analogous to foveal vision in humans: the fovea provides resolution needed for discriminating patterns of

interest, while the periphery provides broad area monitoring for alerting and gaze control.

A foveal strategy that allocates some sensing resources to broad area monitoring, and some to region-of-interest observation can reduce the actual data that needs to be provided by a sensor, and processed by the vision system, by a factor of 1000 or more. This can easily mean the difference between a system that is too large to be considered for any application and one that is sufficiently small to be generally used.

There are two primary areas of research in the area of active vision. The first is in the development of fast, intelligent, control processes to direct the camera. The second is the development of special sensors for foveal vision. Technology for real time control is only beginning to be developed, since real time hardware has been available for only a few years. This work needs to be extended for face location and tracking for the specific application of facial expression recognition. Experimental sensors with foveal organization have recently been built as well. Current devices are too limited in resolution to be considered for practical applications. An alternative is to obtain images at full resolution with a standard camera, then reduce data and resolution electronically to obtain an equivalent foveal sensor. This approach is possible with current technology.

In addition to this work on sensor control, research should be directed to the use of new sensors, such as those that provide range data. Range data has been used in face recognition. The usefulness of such data for recognizing facial expression should be a topic for further study.

Benefits

It is likely that current sensor technology will suffice for the immediate needs of the research community. New 3D sensors could prove very effective. Advanced sensor technology, particularly to control the sensors, will be essential for practical systems for use in medical, computer interface, communication, or other commercial applications.

Detection of Faces

Current status

Discerning the existence and location of a face, and tracking its movements, are perceptual abilities that have not found their own place in the behavioral science literature, but duplicating these native, autonomous functions computationally is not trivial. This task is a precursor to determining the information that the face provides. The strategies that have provided some success in locating faces are described in Tutorial on Neural Networks and Eigenfaces (pages 19 to 21).

Key research challenges

A robust way to locate the faces in images, insensitive to scale, pose, style (with or without eyeglasses or hair), facial expression, and lighting condition, is

still the key research challenge, especially in complex environments with multiple moving objects.

It seems that image segmentation based on the combined use of color, texture, shape (geometry and shading), and model knowledge could provide better performance than most existing algorithms.

For applications that allow careful control of lighting and background, some effort should be directed at designing entire systems and environments for face location and tracking.

Benefits

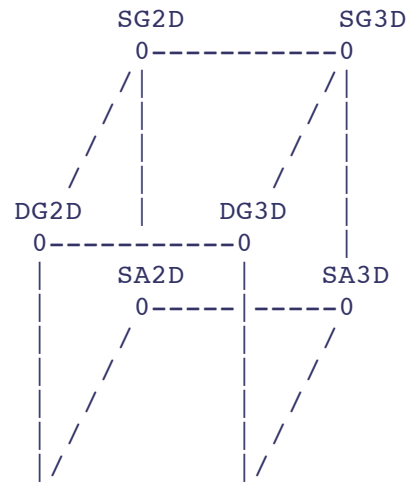
Face detection and tracking are the first steps in face recognition and facial expression understanding. Without knowing where the faces are, most feature extraction algorithms will produce many false targets and thus make themselves less useful. When faces are properly located and tracked, our knowledge about spatial features of a face can be used very effectively.

Feature Extraction from Static Images

Current status

Feature extraction may be divided into at least three dimensions represented in the figure below. The first consideration is static versus dynamic features: Is temporal information (a sequence of images) used or not? The second is the grain of the features: These may be divided into global features, spanning roughly the whole object being analyzed at one extreme, and analytic or part-based features, spanning only subparts of the image. The third is view-based versus volume-based, or 2D versus 3D features. 3D features can be extracted using special sensors or active sensing.

Given this nomenclature, most of computer vision for the last thirty years has been directed towards static, analytic, 2D feature extraction. Optic flow is in the dynamic, analytic, 2D corner. It is of interest here to consider what has been accomplished in these traditional corners, and what might be accomplished in some of the others.



The Necker Cube of image processing. S/D: static/dynamic; G/A:global/analytic; 2D/3D: view-based/volume-based.

Some considerations that have been left out of the above analysis include whether the sensors are active or passive, and whether the features are predefined or learned by an adaptive mechanism depending on the data. A number of computational techniques have been used to extract facial features, and the successes in recognizing a face from a number of faces is contained in the Tutorial on Neural Networks and Eigenfaces (pages 21 to 24. Many of the papers in the references of this section provide reviews of the history and status of this area.

Key research challenges

Most current systems have been applied to small databases; only the eigenface approach has been applied to a large database. It is important to assess these techniques, and contrast them, on large size databases.

The techniques used in recognizing signs of identity should be applied to expression recognition. Some researchers (Cottrell & Metcalfe, 1991; Kohonen et al., 1977; Turk & Pentland, 1991) have looked at expression recognition. Mase, see next section, has looked at optic flow for FACS detection. There is a need to extend these techniques, possibly by principal component analysis, and evaluate them. In order to do this, large, labeled databases are required.

Better techniques are needed to achieve scale invariance and deal with background noise. Combinations of the above techniques could be useful, e.g., one could use eigentemplates for features combined with springs in an energy function approach.

Implementation on parallel hardware could speed up and simplify the algorithms.

These techniques could be used to detect boredom, attention wandering, or sleepiness, but the behavioral scientists have not specified the features required for such performance monitoring. Again, temporal, labeled databases validated by physiological measures are necessary for attacking the problem.

New research should explore different parts of the feature cube diagrammed above. Corners not explored thus far include temporal global features, and the use of 3D part-based features. Can optic flow be used for real-time monitoring? Can active vision techniques reduce the computational requirements of optic flow?

In all of the eigenface/holon approaches, the features extracted were linear. Nonlinear feature extraction through new dimensionality reduction techniques could give lower dimensional representations, and compact parameterizations of expression and face space.

Benefits

Extracting features from faces is the first step in automating a facial expression understanding system.

Feature Extraction from Image Sequences

Basic features: Current status

Changes in the shapes of facial features, their relative positions, and the optical flow in facial areas are parametric features suitable for describing facial expressions. Moreover, they are extractable by computer vision techniques. The possible parametric features are:

- Changes in shape (deformations),
- Changes in positions,
- Flow (Aggarwal & Nandhakumar, 1988),
- Flow of tokens: The extreme points of the lines of the eyes, eyebrows, and lips can be used as tokens and their movements are tracked to extract sparse flow (Sethi & Jain, 1987),
- Flow of areas: Correlation based (Barnard & Thompson, 1980) or gradient based (Horn & Schunck, 1981) optical flow algorithms can extract the flow of regions without using tokens. Exact point correspondence typically does not exist.

Static parametric features have been used in person identification systems (Sakai et al., 1972; Kanade, 1973), and the same algorithms for feature extraction may be worth trying (see pages 15-17, 23-24, and 41).

The face is a good subject for computer vision research, because the shape of facial features and their relative arrangement are universal regardless of age, gender, and race. Consequently we have *a priori* knowledge, and perhaps even a facial model, that can be used to help extract information-bearing features. For instance, standard edge extraction algorithms can, in well illuminated images, detect the eyebrows, eyes, nose (nose wing and nostril) and mouth (upper lip and mouth opening). Yet the lower lip contour and the line of the chin are often not detectable without *a priori* assumptions about the shape and location. Active contour models, such as snakes (Kass et al., 1987; Waite & Welsh, 1990) and deformable templates (Yuille et al., 1989), are one way to accomplish this. After extracting the features, their positions, such as centroid, extreme point, shape and angle, are used to analyze and/or to represent the expression (Choi et al., 1990; Terzopoulos & Waters, 1990a).

In contrast, the optical flow approach (Mase & Pentland, 1990a, 1991; Mase, 1991) to describing face motion has the advantage of not requiring a feature detection stage of processing (see pages 24-25).

Key challenges

Robustness to temporal noise. All of the above parameters except optical flow are theoretically computable from static imagery. Such extraction has, however, proven to be sensitive to noise and illumination. Optical flow information is also affected by these problems, however spatial averaging over facial action groups

seems to offer some hope of robust estimation.

Skin deformations. An important but often unavoidable source of noise during expressions is the appearance of wrinkles and dimples. They are confusing for feature extraction techniques, and violate the constant-patch assumption of the optical flow computation. It is necessary for all algorithms to be able to deal with these "irrelevant" features by, for instance, separating these features in terms of their temporal stability.

Head motion. Before one can interpret the detailed motion of the face, it is first necessary to very accurately track the head. This subject is discussed in more detail in previous sections (see pages 21-22 and 39-41).

Temporal segmentation. Temporal segmentation is necessary for a system to pull out each separate expression from within a long sequence of facial actions; this is particularly true if we are to understand dialogs between people. In lip reading, zeros of the velocity of the facial motion parameters were found to be useful for the temporal segmentation (Mase & Pentland, 1990b). This finding may be useful in attempting to segment more complex facial actions.

Co-articulation. It is well known in speech recognition that adjacent sounds are often co-articulated; that is, the temporal context of a phoneme changes its sound. Facial expression seems to be similar. Analysis of the mechanisms of co-articulation and compensating for them in the recognition stage is a major challenge (Garcia et al., 1992).

Higher level feature extraction. The basic information, such as shape deformation, position change, and optical flow, may be integrated spatially to obtain higher level descriptions, such as muscle actions and Action Unit (AU) descriptions. The use of anatomical knowledge is necessary in this task; however careful statistical analysis has also proven to be useful (Mase & Pentland, 1990b; Mase, 1991).

Benefits

Feature extraction and computation of facial changes is likely to be the basis for accurate expression description. Description of facial expression based on well-defined and well-segmented information will lead to a reliable recognition of expression.

Lip Reading

Computerized lipreading: Background

A main objective in trying to elicit spoken language from optical observations of the oral cavity and facial articulatory movements is to supplement the acoustic perception of speech, particularly important in noisy environments. Also, understanding the relations between speech and observable articulation is useful in teaching the art of lipreading to the hearing-impaired and in the simulated animation of synthesized speaking faces.

Automating the recognition of speech from video signals - also called Optical

Automatic Speech Recognition or OASR - is expected to significantly enhance the robustness of acoustic automatic speech recognition because of the complementarity between ambiguous phonetic and visual signals. In the case of human lipreading the experimental gains are on the order of 10 to 12 db improvement in the signal-to-noise ratio, according to Brooke (1989). Brooke also suggests that non-articulatory measurements of facial motions (head, eyebrows, etc.) may be auxiliary to augmentation of acoustic recognition in non-ideal circumstances by providing nonverbal cues.

The earliest reported attempt to mechanically automate lipreading is the patent number 3192321 issued in 1965 to Ernie Nassimbene (1965) of IBM for a device consisting of an array of photocells that captured the reflected light emitted from the oral cavity region. Subsequent research by Petajan, Brooke, Nishida, Pentland, Mase, Smith, Yuhas and others is described in the Tutorial Neural Networks and Eigenfaces on pages 25 to 26.

Mouth region features for lipreading

The work of Garcia and Goldschen (Garcia et al., 1992) in analyzing the features that are most important for continuous speech recognition using TIMIT sentences is described on page 25. Previous work by Montgomery and Jackson (1983) had sought to elicit important features for lipreading vowels in cases where /h/ precedes each vowel, and /g/ follows each vowel. The features examined were the height, width, area, and spreading (width/height) of the oral cavity, video duration (number of frames for vowel articulation), and audio duration (time during vowel articulation using an oscilloscope). These features were taken from a single image frame chosen by experienced lip-readers to characterize the vowel. They concluded that the spreading and the area surrounded by the lips are important features for vowel recognition. More significantly, they found no absolute fixed and observable positions for the lip and tongue corresponding to specific vowels across different speakers. Lip-readers adjust to different speakers through spatial normalization. That is, constant relative differences among some oral cavity features in the visual perception of the vowel utterances exist for all speakers.

Kathleen Finn (1986) investigated appropriate oral cavity features for possible automatic optical recognition of consonant phonemes. Finn investigated the recognition of consonants preceded by and followed by the vowel /a/. Her analysis considered only the middle optical image frame of the utterance for each consonant, thereby taking into account only central static features for viseme discrimination. Finn determined that the five most important static features for consonant discrimination are the height and width of the oral cavity opening, the vertical spreading (each separately) of the upper and lower lips, and the "cornering" of the lips.

Research issues in OASR

One of the most important issues in OASR is how to automate the recognition process in real time. The use of powerful workstations makes possible real time acoustic recognition. As continuous progress on real-time computer vision systems is anticipated, we expect supplementary optical speech recognition

also to make comparable progress in real-time computational environments with large vocabularies. A few commercial systems for acoustic speech recognition are available at the present time, and other academic research systems have even shown speaker-independent performance (Lee, 1989). One of the challenges involving human/computer interfacing is how to improve the robustness of recognition using a multi-modal approach. The approach taken by Garcia and Goldschen is to use a Hidden Markov Model (HMM) that parallels the acoustic model, which can therefore be augmented with the additional features obtained from a video camera.

An alternative approach is to use time-dependent neural networks that would allow complex utterances to be recognized beyond the phonetic boundaries of a single phone, providing co-articulation information. Given the difficulty of training and recognizing a large number of different context-dependent phones, it seems that the association of optical speech recognition using neural nets with acoustic recognition using neural nets, for large vocabularies or continuous speech, must await for further research developments. For small vocabularies, Stork, Wolff and Levine (1992) showed the robustness of recognizing both optical and acoustic signals over a purely acoustic recognizer.

The problems of phonetic context dependency, which plague the acoustic recognition of speech, also appear in the optical channel. The mapping between phonemes and visemes has been an area open to argumentation, clearly because context dependency obscures fundamental issues between the actual phones and their correspondingly observed sequence of optical images. Solution of the problems of co-articulation is a likely prerequisite for continuous optical recognition, as contrasted with isolated word recognition.

Another, more fundamental issue, is to what extent optical features can be considered speaker independent, and whether the training techniques for speaker independent acoustic speech recognition are also applicable to optical speech recognition.

Lip synchronization

The inverse problem to analysis of facial movements in the mouth area -- having the objective of speech recognition -- is the synthesis of the facial motions that take place, given spoken utterances. This problem has obvious commercial implication for computer animation of "talking heads."

Expression Recognition

A final automated expression recognition system must translate the automatically extracted features into a description of facial expression. It is usually expected that this automatic description should be identical, or at least very close to, a human's description of a facial expression. But in general, the requirements of expression recognition will depend on the applications, so the isolation of action units may require a much finer resolution than a simple classification, such as between sad or happy.

Current status

Our present experience on expression recognition is still limited. Usually the work is restricted to frontal images of faces under good illumination. Optical flow has been used to extract dynamical muscle actions. These actions formed a fifteen-dimensional feature vector, which was categorized into four expressions using a nearest neighbor technique (Mase, 1991). The eigenface approach has also been used to successfully classify expressions for a single person (Turk & Pentland, 1991). More recently, a variation in the eigenface approach has successfully classified the six basic emotional states across a database of eleven people (Pentland et al., 1992).

More work has been done on face recognition, including gender classification. Many face recognition systems skip the feature extraction step completely and solve the recognition problem by template matching of the test image with all target images, or in other words each face is its own feature. To reduce the amount of computation, often a smaller set of images is used as templates. Then, the features are the coefficients of the linear representation of the test image using these templates.

Some other nonlinear approximation techniques, such as Hyper-basis functions, may help to get a better understanding of the importance and role of certain features for expression recognition. Using this method, it was found that the position of the eyebrow is more important than the position of the mouth for gender classification (Brunelli & Poggio, 1991). However, Golomb et al. (1991) and Gray et al. (1993) have found that the shading around the filtrum and mouth area provide significant information about sex-identity.

Key research challenges

The main problem in this area is generalizing from given examples of expressions to ones the recognition system has never seen before. It is important that the given examples should be described not only as static images of faces but also as dynamical sequences of images, as in many cases the expression of a face is determined by the temporal changes in the face as much, as by the final shape. We expect to obtain this precise description from psychology and physiology.

Generalization or learning from examples is equivalent to function approximation in higher dimensional spaces (Poggio, 1990). We want to find the functional dependence between the input space, the space of the different features and the output space - the space of different expressions. The success of such learning depends on the quality of the extracted features. If, for example, the features are already viewpoint or illumination independent, the number of necessary examples will decrease.

Unsupervised learning is also possible. In this case, one would try to find some characteristic differences in the extracted feature sets and use these differences afterwards for a classification. Most techniques in this area fit in the projection pursuit framework (Friedman & Stuetzle, 1981). The development of these classes is not guided by human understanding of expression. The success of this method, a correlation between the evaluated classes and our understanding of different expressions, will again depend on the type of

extracted features.

NOTE: This section was prepared by A. Yuille and A. Pentland, from contributions of their own and P. Burt, G. Cottrell, O. Garcia, H. Lee, K. Mase, and T. Vetter, and edited by T. S. Huang.

IV-C. Breakout group on Modeling and Databases

Participants: F. Parke, D. Terzopoulos, T. Sejnowski, P. Stucki, L. Williams, D. Ballard, L. Sadler, J. Hager, D. Psaltis, and J. Zhang

Computer-Based Facial Expression Models and Image Databases

Fred Parke, Demetri Terzopoulos, Terrence Sejnowski, Peter Stucki, Lance Williams, Dana Ballard, Lewis Sadler, and Joseph Hager

Introduction

The ability to recognize and generate animated facial images, together with speech input and output, holds enormous promise for many diverse application areas. Consider the following examples.

The use of model based scene analysis and model based facial image synthesis could yield very low bandwidth video conferencing (Aizawa et al., 1989; Choi el al., 1991).

Applications relevant to human/computer interface development include

- lipreading systems (e.g., AT&T Bell Laboratories, Murray Hill, NJ 07974),
- gaze tracking devices (e.g. RK-426 pupil/corneal reflection tracking system, Iscan, Inc., 125 Cambridge Park Dr., Cambridge MA),
- ultrasonic head-tracking for steering the zones of an autostereo display (Dimension Technology, Inc., 176 Anderson Ave., Rochester, NY 14607),
- and computer animation of talking heads as a supplement to text-to-speech audio (e.g. British Telecom Research Laboratories [Welsh et al., no date] -- commercial software includes "InterFace" and "At Your Service," by Bright Star Technologies, 325 118th SE, Bellevue WA 98005).

Substantive research in the real-time animation of faces for telecommunication and for the synthesis of computer interface "agents" is being conducted at Apple Computer, Inc. (Advanced Technology Group, MS:76-4J, 20525 Mariani Ave., Cupertino, CA 95014), Hitachi, Ltd. (Hitachi Central Research Laboratory, 1-280 Higashi-Koigakubo, Kokubunji, Tokyo 185, Japan), NTT (Human and Multimedia Laboratory, 420 C, NTT Human Interface Laboratories, 1-2356 Take, Yokosuka-Shi, Kanagawa, 238-03 Japan), and Sony (Information Systems Research Center, Sony Corporation, Asahi-cho, Atsugi-shi 243, Japan).

A number of companies are in the business of vending computer systems and services for making facial image composites ("identikit" police identification tools, point-of-purchase video preview for cosmetic make overs or cosmetic surgery, and one class of systems for estimating the aged appearances of missing children), 3D digitization of faces, 3D reconstructive surgery preview and manufacture of facial prosthetics, 3D digitization of teeth for the manufacture of dental appliances, and 2D and 3D facial animation.

Another important current interest is in the entertainment industry; the use of graphical face models in advertising, for movie special effects, etc.

These many diverse examples illustrate the potential of a focused research program directed towards computer understanding of facial expression. The purpose of this document is to delimit the specific research questions that would form the basis of such a program. To do this, the document is organized into four subsequent sections: 1) 3D Modeling, 2) Facial Databases and Security, and 3) Research Directions.

State of the Art in 3D Modeling

Brief partial history

The first work in developing facial models was done in the early 70's by Parke at the University of Utah (Parke, 1972a, 1972b, 1974, 1975) and Gillenson at Ohio State (Gillenson, 1974). Parke developed the first interpolated and the first parametric three dimensional face models while Gillenson developed the first interactive two dimensional face models. In 1971, Chernoff (1971, 1973) proposed the use of simple 2D computer generated facial images to present n-dimensional data. In the early 80's, Platt and Badler at the University of Pennsylvania developed the first muscle action based facial models (Platt, 1980, 1985; Platt & Badler, 1981). These models were the first to make use of the Facial Action Coding System (Ekman & Friesen, 1978; Ekman & Oster, 1979) as the basis for facial expression control.

The last seven years has seen considerable activity in the development of facial models and related techniques. Waters and Terzopoulos developed a series of physically based pseudo-muscle driven facial models (Waters, 1986, 1987, 1988; Waters & Terzopoulos, 1990, 1992; Terzopoulos & Waters, 1990b). Magnenat-Thalmann, Primeau, and Thalmann (1988) presented their work on Abstract Muscle Action models in the same year as Nahas, Huitric and Sanintourens (1988) developed a face model using B-spline surfaces rather than the more common polygonal surfaces. Waite (1989) and Patel and Willis (1991) have also reported recent facial model work. Techniques for modeling and rendering hair have been the focus of much recent work (Yamana & Suenaga, 1987; Watanabe & Suenaga, 1992). Also, surface texture mapping techniques to achieve more realistic images have been incorporated in facial models (Oka et al., 1987; Williams, 1990; Waters & Terzopoulos, 1991).

The ability to synchronize facial actions with speech was first demonstrated by Parke in 1974 (Parke, 1974, 1975). Several other researchers have reported work in speech animation (Pearce et al., 1986; Lewis & Parke, 1987; Hill et al.,

1988; Wyvill, 1989). Pelachaud has reported on recent work incorporating co-articulation into facial animation (Pelachaud, 1991). Work modeling the physical properties of human skin have been reported by Komatsu (1988), Larrabee (1986), and Pieper (1989, 1991).

Current models

Essentially all of the current face models produce rendered images based on polygonal surfaces. Some of the models make use of surface texture mapping to increase realism. The facial surfaces are controlled and manipulated using one of three basic techniques: 3D surface interpolation, *ad hoc* surface shape parameterization, and physically based with pseudo-muscles.

By far the most common technique is to control facial expression using simple 3D shape interpolation. This is done by measuring (Cyberware Laboratory Inc., 1990; Vannier et al., 1991) the desired face in several different expressions and interpolating the surface vertex values to go from one expression to the next. One extension on this approach is to divide the face into regions and interpolate each region independently (Kleiser, 1989).

Ad hoc parameterized facial models have been developed primarily by Parke (1982). These models present the user with a small set of control parameters that manipulate various aspects of facial expression and facial conformation. These parameters are only loosely physically based. These parametric models are the only ones to date that allow facial conformation control, i.e., changes from one individual face to another.

Physically based models attempt to model the shape changes of the face by modeling the properties of facial tissue and muscle actions. Most of these models are based on spring meshes or spring lattices with muscle actions approximated by various force functions. These models often use subsets of the FACS system to specify the muscle actions.

Deficiencies

Even the best current physically based facial models use relatively crude approximations to the true anatomy of the face. The detailed structure of the face (Fried, 1976; Friedman, 1970), its anatomical components, their interaction, and actions are only approximated at fairly abstract levels.

No model to date includes the complete set of FACS muscle actions. The muscle action models do not yet model the true muscle behavior, but only approximate it with various fairly simple mathematical functions

Facial Databases

Throughout this report, repeated references to the need for a database of facial information have emphasized the glaring lack of such a resource. The preceding sections have indicated many types of information that should be part of such a database, including images of the face (both still and motion), FACS scores, interpretations of facial signs, physiological data, digitized speech

and sounds, and synthetic images. Tools for manipulating images and associated data and for synthesizing facial images might usefully be linked to the database. Much of the progress in applying computers to understand the face depends upon access to a facial database, both in terms of sharing information and common sets of images, and in regard to increasing cooperation and collaboration among investigators.

The benefits of the image database include reduction of research costs and improved research efficiency. Costs can be reduced by avoiding redundant collection of facial expression exemplars by each investigator independently. The retrieval and exchange of images is also improved by a centralized repository. Research efficiency can increase by maintaining a reference set of images that can provide a basis for benchmarks for efforts in various areas. Also, a great deal of information will be collected about this reference set.

Because the image part of the database is clearly the first step in constructing a full, multimedia database of the face, it is discussed more fully in the following section on images and security.

Images in the database

Several technical considerations for database images need to be resolved, including criteria for image resolution, color, and sequence, data formats, compression methods, and distribution mechanisms. The choice should be compatible with with other multimedia, scientific databases now being developed in biology and medicine.

Images for the database must meet the multiple needs of scientists working in this field, independent of the technical criteria for inclusion in the database. In other words, some images of importance may not meet current standards for resolution or color. These images probably include those from the archives of leading behavioral scientists. An important aspect of a database of images is the metadata associated with each image, such as the circumstances under which the image was obtained. Some metadata and other associated fields are discussed below.

For general relevance, the images should be scored in terms of Ekman and Friesen's FACS (1978) as a standard measure of activity in the face. Images should represent a number of demographic variables to provide a basis for generality of research findings. These variables include racial and ethnic background, gender, and age. For each category of images in the database, images of several individuals need to be included to avoid effects of the unique properties of particular people.

Another important variable, in terms of both signal value and neurophysiology, is the distinction between deliberate actions performed on request versus spontaneous actions not under volitional control. Many expressions are not easily classified into either of these categories. Examples of both categories (with provision for additional categories) must be included in the database in order to study the substantive question of the difference between these expressions. Examples of the prototype expressions for each of the seven or

eight basic emotions should be included. In addition, expressions in which only part of the prototype is present, in one or more areas of the face, should be available. Examples of deliberate individual muscular actions and the important combinations of actions for selected experts in facial movement is an important component of this database. Intensity of muscular action should be varied for many of the image categories above.

For spontaneous expressions, an important consideration is to identify within the database the precise nature of the conditions that elicited the expression. The eliciting circumstances, such as a conversation versus watching films alone, can produce different types of expressions.

Given the number of classifications and variables, the number of images to fill a full matrix of cross-tabulations would be quite large. The great number of expressions that might be included is one reason for including a module in the database that could artificially generate variations on expressions given certain image prototypes. Such images could be used as stimuli in judgment studies.

Both still images and motion images of the facial expressions are required in the database. Still images can convey much of the configurational information about facial expressions that is important for inferring meaning in terms of emotion, message, etc. Motion records contain information about the temporal dynamics of expression that might convey information about the volitional quality of the movement, deception in the expression, etc. Both still and motion records could provide the basis for computer-based measurement of the image. Motion records would be important for work on animation and modeling.

In addition to the reference images installed and managed by the database administrator, provision should be made for individual researchers to add their own images to the database. These additions must match the specified technical formats which should be made available to all researchers for their equipment purchase and record collection phases of research. The security status of such additions also needs to be determined.

In order to meet the high resolution technical requirements of the database, most images used as references in the database will need to be collected from scratch. A limited number of images are currently available that might be included in the database to start. An indication of where such archives might be is contained in the database recommendations of the Basic Science Workgroup on page 35. A survey should be made of these and other laboratories to determine what images exist that might be contributed to the database, what restrictions they have, and the terms under which they might be incorporated. Color moving images of spontaneous facial behaviors will probably be difficult to find on anything but consumer level video in these archives.

Security Issues

The database of images can be valuable to thousands of researchers if it is easy to access and use. Ease of access implies a relaxed level of security that allows most users quick access to materials they need and frees the administrator of time consuming identity checks. In this case, the database

could be accessed by nonscientists, such as journalists, advertisers, and ubiquitous hackers. However, some investigators may need to work with images that can be made available only to certain authorized users. Examples include images of psychiatric patients with emotional disorders or facial surgery patients who are unable or unwilling to give permission to use their likenesses outside a community of scientists. These images could easily, even unwittingly, be abused by users who are not familiar with informed consent procedures, granting agencies's regulations on uses of identifiable records, or permissions. Such examples indicate the need for a more comprehensive security strategy.

Issues related to security are very important in any database design. Their requirements, however, vary from business to engineering to scientific databases. While these requirements are rather well understood in the business and engineering databases, the problem of security has not yet been addressed in full detail for the design of scientific databases.

For the subject image database, there are many aspects that need to be investigated by an interdisciplinary team of database designers, computer vision experts and behavioral scientists. For example, in the area of access control, the following model-options are possible:

a) The Discretionary Access Control (DAC) model, where the owner of objects or object classes keeps all rights and can grant or revoke them to individual users or user-groups of his choice. The granting can include units like tuples or sub-units like attributes. Also, the database end-user community can get the right to use a complex object like a human face but not the right to use details or underlying contraction/expansion models of the same human face. The database manager has no direct influence on this security scheme.

b) Mandatory, Multilevel Access Control (MAC) model. The users can get rights to use individual objects or object classes at various security levels (registered confidential, confidential, for internal use only, no security). The assignment of security access control is rule-based and can be implemented by the database manager.

Both access control models are based on proper identification and authentication by the database user or user-groups.

A MAC model for secure use of databases might be implemented as a three tiered hierarchy. In the lowest security level, all images should be without restriction on use, i.e., "in the public domain." This criterion would exclude any images that have proprietary restrictions (e.g. use with royalty), have any restrictions on their use (e.g., consents or permissions), or are of a sensitive nature (e.g., patients or prisoners). This level of security allows anonymous or guest access to the database. The medium level of security includes images that can be used only for scientific purposes, without restrictions such as royalties. All images at this level should be accessible by anyone who is authorized at this level. The high level of security gives access to images that have special considerations requiring careful validation of users. This level of security might be used by police, security, and military agencies, by those who require some kind of payment for use of images, and by researchers who have

sensitive images that require screening of users. The high level of security should be flexible enough for a number of different categories of users, each with their own databases. The medium and high levels of security imply a system of userids and passwords and an administrative staff to verify user identities and maintain the user database.

Most currently available images would seem to fall into the medium level of security, but the status of existing images needs to be investigated. Procedures for determining the appropriate security level need to be developed and the role of the owner or contributor in this process specified. Incentives for obtaining images for lower levels of security from primary researchers would be helpful.

Another import issue is the problem of object distribution over networks. This requires appropriate network security level up to encrypted procedures. Again, the assignment of distribution security can be done on an object- or sub-object level.

Research in database design has gradually evolved from business databases to engineering databases and is currently starting to address issues concerned with scientific databases (Frenkel, 1994; Rhiner & Stucki, 1992). Similar security issues arise in the human genome project.

Research Directions and Goals

Anatomically correct models Models which faithfully reflect detailed facial anatomy, facial muscle actions, and facial tissue properties need to be developed. These models would be useful both for medical applications and for expression understanding research. The Library of Medicine is creating a three-dimensional model of several human bodies from digitized slices. The face portion of this database could provide an anatomically accurate model for the facial musculature.

Expression control Models which include complete, accurate expression control capabilities are needed. Control could be based on the FACS system or other parameterizations as needed and developed. The ability to easily specify and control facial conformation (those facial aspects which make each face unique) is desirable. Orthogonality between expression control and conformation control is essential.

Universal models An ultimate goal is the development of models which quickly and accurately provide representations for any face with any expression. This goal should address the extent to which artificially generated images can be substituted for photographic images.

Accessible models These models needed to be readily available (in the public domain?) and supported on common platforms - PC's and workstations.

Database organization A database of images needs answers to such questions as: How does one construct and administer large multimedia databases? What is the relationship between performance and database organization of the multimedia database? How do you distribute images and other fields efficiently

and quickly via network or storage device? How do you search for key properties of a visual image?

Database security issues As discussed above security of databases can potentially be a huge problem. Some method of managing personal information must be established that has different levels of protection. The Human Subjects protocol used by universities could serve as a model.

NOTE: This report was assembled by D. Ballard from his own contribution and others from F. Parke, J. Hager, L. Sadler, P. Stucki, D. Terzopoulos, T. Sejnowski, and L. Williams, and was edited by T. Sejnowski.

V. RECOMMENDATIONS

Joseph C. Hager

The recommendations of this Workshop are summarized here under four major headings: Basic Research, Infrastructure, Tools, and Training. These recommendations were collected from the rest of this report, where justification may be found, and from the proceedings of the Workshop. The list is long, emphasizing the specific, important research and development efforts needed to advance this area. The list should help make concrete the extensive nature of the tasks, although it is not exhaustive of worthwhile contributions. By focussing on specifics, rather than general efforts, more precise guidance is provided to investigators about the research questions and to funding institutions about required support.

Basic Research on the Face that Answers These Crucial Questions:

- A. In regard to the system of facial signs and their perception:
  - i. How much of a known prototypical emotion expression is needed for observers to recognize it?
  - ii. Are there differences among specific cultures in perception of basic emotion expressions?
  - iii. What variations on the prototypical emotion expressions are still perceived as belonging to the same basic emotion?
  - iv. How are expressions that contain only some of the actions of the prototypical expression judged?
  - v. How are blends of different emotion expressions within the same expression perceived and judged?
  - vi. What is the effect of asymmetry on the perception of emotion expression and other signs?
  - vii. How does the perception of emotion expression affect the perception of the expresser's other personal characteristics?
  - viii. What about an expression indicates deception versus honesty, genuine involuntary expression versus feigned or deliberately produced expression, and suppression or augmentation of an involuntary expression?
  - ix. How do emotion signs interact with other facial signs to produce an interpretation by an observer?
  - x. What is the relative contribution of the face compared to other signal systems, such as the body and language, and how are these systems integrated by the expresser and decoded by the observer?
  - xi. What are the temporal dynamics of facial movements and expressions and does this provide additional information beyond the configuration?

        - xii. What effect do co-occurring or adjacent muscular movements in the sequence of facial movements have on each other (co-articulation)?
    - B. What are the subjective and physiological consequences of voluntary production of partial and complete prototype emotion expressions?
    - C. In regard to spontaneous expressive behavior:
        - i. What is the range and variability of spontaneous facial expressive behavior?
        - ii. How frequent are the facial prototypes of emotion in real-world situations?
        - iii. Across cultures, what components of spontaneous emotion expressions are invariant within an emotion family?
        - iv. How do expressive behaviors, such as head and lip movements, contribute to the perception and interpretation of facial expressions?
        - v. What are the relationships among speech behaviors, vocal expression, and facial expressions?
        - vi. What information is carried in physiological indices that are not available in expressive measures?
    - D. What behavioral and physiological indices differentiate between voluntary and spontaneous expressive productions?
    - E. What facial behaviors are relevant to human performance and how can these be monitored automatically?
    - F. In regard to the neural and physiological basis for facial messages:
        - i. What are the neural correlates of the perception of facial information specifically the neural centers for facial perception, their structure, and their connections to other neural centers?
        - ii. What are the neural correlates for the production of facial expression, specifically the neural centers for involuntary motor action, their efferent pathways, and their connections to other effector neural centers?
        - iii. What, if any, are the relationships between the production and the perception of facial expressions in regard to neural activity?

Infrastructure Resources that Include the Following:

- A. A multimedia database shared on the Internet that contains images, sounds, numeric data, text, and various tools relevant to facial expression and its understanding. The images should include:
    - i. still and moving images of the face that reflect a number of important variables and parameters, separately and in combination,
    - ii. voluntary productions of facial movements, with descriptive tags, accompanied by data from other sensors,
    - iii. spontaneous movements carefully cataloged with associated physiology,
    - iv. animation exemplars for use in perception studies,
    - v. compilation of extant findings with more fine-grained description of facial behavior,
    - vi. large numbers of standardized facial images used to evaluate performance of alternative techniques of machine measurement and modeling.
    - vii. speech sounds and other vocalizations associated with facial

messages.
- B. A survey of what images meeting the criteria in A. above currently exist and can be incorporated into the database, and a report of the images that need to be collected.
- C. Specifications for video recordings and equipment that would enable sharing the productions of different laboratories and that anticipate the rapid developments in imaging and image compression technology.
- D. Standards for digitized images and sounds, data formats, and other elements shared in the database that are compatible with other national databases.
- E. A security system that protects the privileged nature of some items in the database while maximizing free access to open items.
- F. Analysis of database performance and design.
- G. Strategies and opportunities to share expensive equipment or complex software among laboratories.

Tools for Processing and Analyzing Faces and Related Data:

- A. Methods for detecting and tracking faces and heads in complex images.
- B. Programs to translate among different visual facial measurement methods.
- C. Automated facial measurements:
  - i. detecting and tracking 3D head position and orientation,
  - ii. detecting and tracking eye movements, gaze direction and eye closure,
  - iii. detecting and measuring lip movement and mouth opening,
  - iv. detecting and measuring facial muscular actions, including the following independent capabilities:
    - a. detection of brow movements,
    - b. detection of smiles,
    - c. detection of actions that are neither smiling or brow movements,
    - d. techniques for temporally segmenting the flow of facial behavior,
    - e. detection of onset, apex, and offset of facial muscle activity,
    - f. detection of limited subsets of facial actions.
- D. Parametric and other models, including 3D, of the human face and head that enable accurate rendering of different expressions given a specific face.
  - i. anatomically correct physical models of the head and face,
  - ii. complete image atlas of the head, including soft and hard tissue.
- E. Algorithms for assembling discrete measurements into meaningful chunks for interpretation.
- F. Programs for translating facial measurements in terms of emotion, cognitive process, and other phenomena incapable of direct observation.
- G. Programs for translating lip movements to speech.
- H. Automated gesture recognition.
- I. Programs for integrating and analyzing measurements of different modalities, such as visual, speech, and EMG.
- J. Pattern discovery and recognition in multiple physiological measures.

- K. Further exploration of novel computer vision and image processing techniques in processing the face, such as the use of color and 3D.
- L. Development of "real-time" distance range sensors useful in constructing 3D head and face models.
- M. Development of interactive systems for facial analysis.
- N. Development and adaptation of parallel processing hardware to automated measurement.
- O. Video sensors and control equipment to enable "active vision" cameras that would free behavioral scientists from requirements to keep subjects relatively stationary.

Training and Education for Experienced and Beginning Investigators:

- A. Resources providing specialized training:
  - i. Post-doctoral, interdisciplinary training using multiple institutions,
  - i. Summer Institutes for the study of machine understanding of the face,
  - ii. Centers of excellence in geographic centers where high concentrations of relevant investigators and resources exist.
- B. Resources to facilitate communication among investigators:
  - i. Special journal sections to bring information from different disciplines to the attention of other relevant disciplines,
  - ii. Computer bulletin boards,
  - iii. On-line journal or newsletter publishing and information exchange,
  - iv. Liaison to the business/industrial private sector.

NOTE: These recommendations were compiled by J. Hager from the discussions, reports, and working sessions of the Workshop.

VI. BENEFITS

Beatrice Golomb and Terrence J. Sejnowski

The potential benefits from efforts to understand the face using the technologies discussed in this report are varied and numerous. The preceeding sections have separately enumerated many benefits. This section summarizes these benefits and indicates additional areas where benefits could accrue that were not emphasized previously.

Commercial Applications

Automated systems that process natural facial signals and/or generate synthetic outputs related to the face have important commercial potential, as indicated throughout this report. Some products would become parts of systems used by the mass consumer market. For example, machines that monitor the face could become part of digital video conferencing and video telephone systems of low bandwidth, tracking both the location and signals of the face. When combined with synthesis of the face and speech recognition and translation, these systems could become universal translators where a foreign language and the correct lip movements and facial expressions are imposed on audio-video images of a speaker. These "talking heads" could significantly enhance international commercial and political communications. Another area of application is the "personal agent" of advanced personal computer operating systems, enabling users to create and communicate with a customized personality that could perform increasingly numerous and complex tasks of the computer and assess the reactions of the user. This agent would use speech recognition, possibly supplemented with lip reading, and nonverbal facial cues to interpret the user. The agent's face would be created synthetically and would respond with synthesized speech and artificial expressive movements based on the methods described by the Modeling and Database Workgroup (pages 48-50).

Other markets for this technology include specialized areas in industrial and professional sectors. For example, monitoring and interpreting facial signals are important to lawyers, the police, and intelligence or security agents, who are often interested in issues concerning deception or attitude. Machine monitoring could introduce a valuable tool in these situations where only informal interpretations are now used. A mechanism for assessing boredom or inattention could be of considerable value in workplace situations where attention to a crucial, but perhaps tedious task is essential, such as air traffic control, space flight operations, or watching a display monitor for rare events. A special advantage of machine monitoring in these situations is that other people need not be watching, as in the Big Brother scenario; instead, "one's own machine" provides helpful prompts for better performance. Developments in

commercial motion picture production have already taken advantage of digital image processing (morphing, etc.) and would be a likely benefactor of improved synthesis of human qualities. Also in the entertainment area, the nascent virtual reality field might use similar resources.

Finally, the specialized market for scientific instrumentation should not be overlooked. Packaged tools that analyze or synthesize facial information could be a profitable enterprise for researchers in fields that are discussed below.

Computer Sciences

The problem of understanding the face will enhance the computer sciences and technologies brought to bear on these issues. As the report from the Sensing and Processing Workgroup points out (page 43), the face has some ideal qualities for development of computer vision systems. The application of neural networks and innovative approaches to digital image processing is sure to be applied to other problems, such as analysis of planetary photos and particle decay images.

The Modeling and Database Workgroup emphasized the need for a database of the face. There are parallels between the need to collect and maintain image databases for the analysis of facial expressions and the need to develop a database of utterances for speech recognition. The TIMIT database, a joint effort of Texas Instruments and MIT, provided the speech community with a research standard for speech records and a means for objectively comparing speech recognition systems. The proposed database of facial expressions could provide similar benefits for the vision community. As the Database Workgroup indicates, there are many challenges that the creation of such a huge multimedia database would bring to the investigators developing scientific databases.

Basic Science Research

Basic research that uses measures of the face and facial behavior would reap substantial benefits from inexpensive, reliable, and rapid facial measurement tools. In brief, such tools would revolutionize these fields by raising the quality of research in which reliability and precision currently are nagging problems, by shortening the time to conduct research that is now lengthy and laborious, and by enabling many more researchers, who are presently inhibited by its expense and complexity, to use facial measurement. More studies of greater quality at lower cost in more scientific areas is the promise of success in automating facial measurement.

One can imagine the impact of these innovations by considering the breadth of some of the current research that uses facial measurement. In behavioral science, facial expression is an important variable for a large number of studies on human interaction and communication (e.g., Ekman et al., 1972; Ekman & Oster, 1979); is a focus of research on emotion (e.g., Ekman, 1984), cognition (e.g., Zajonc, 1984), and the development of infants and young children, (e.g., Camras, 1977); and has become a measure frequently used in psychophysiolgical studies (e.g,, Davidson, et al., 1990). In anthropology, the

cross-cultural perception and production of facial expression is a topic of considerable interest (e.g., Ekman & Friesen, 1986). For political science and economics, measurement of facial expression is important in studies of negotiations and interpersonal influence (e.g., McHugo et al., 1985). In neurophysiology, correlates of viewing faces, and in some cases facial expressions, with single neuron activity helps map brain function, such as the cells that respond selectively to faces (some to identity, others to expression) in the superior temporal sulcus (Perret et al., 1982, 1984; Rolls et al., 1987, 1989; Baylis et al., 1985), in parietal and frontal regions (Piagarev et al., 1979), and the inferotemporal region (Gross et al., 1972). The amygdala, where neurons also respond to faces, appears to be concerned with emotional and social responses, and ablation leads to inappropriate social responses to faces (Rolls 1984). In linguistics, coarticulation strategies in lower lip protrusion movements (Perkell, 1986), relations between facial nerve, facial muscles, and soft palate in speech (Van Gelder & Van Gelder 1990), features of acquired dysarthria in childhood (Van Dongen et al., 1987), and lip-reading as a supplement to auditory cues in speech perception have been investigated. The Basic Science Workgroup outlined many tools (pages 32 to 38) that could be applied to enhance these efforts.

Medicine

This section examines more closely one area where the facial measurement tools described above could expand and enhance research and applications. Faces and facial expression have relevance to medicine, neurology, and psychiatry, and a system to automate coding of facial expression would advance research in diverse domains.

Many disorders in medicine, particularly neurology and psychiatry, involve aberrations in expression, perception, or interpretation of facial action. Coding of facial action is thus necessary to assess the effect of the primary disorder, to better understand the disorder, and to devise strategies to overcome the limitations imposed by the disorder. In addition, because different disorders produce different effects on expression, examination of facial action skills (production and reception) may assist diagnosis.

In the psychiatric domain, the ability to produce or interpret facial expression is selectively affected by certain brain lesions or psychopathology. Schizophrenia and "psychosomatic" illness lead to blunting of expression, both in patients and in control subjects talking to these patients, who do not consciously know they are talking to mentally ill people (Krause et al., 1989; Steimer-Krause et al., 1990). Facial expression of emotion distinguishes depressed patients on admission to the hospital and after successful therapy (Ellgring, 1989), and various studies have examined expression with major affective illness (Ekman & Fridlund, 1987) and unipolar depression (Jaeger et al., 1986), as well as other psychopathology (Mandal & Palchoudhury, 1986). Automated methods for assessing facial responses and delivering stimuli would improve the research on and delivery of clinical services.

In neurology, analysis of inappropriate facial expressions may provide evidence for the location and type of brain lesions. Brainstem damage may lead to

emotional lability, as in pseudo-bulbar palsy. Changes of facial expression consistent with sadness, fear, surprise, etc. have also been described at the onset of seizures (Hurwitz et al., 1985), and outbursts of anger have been seen with brain lesions as well (Poeck, 1969; Reeves & Plum, 1969). Parkinson's disease, a disorder of the dopaminergic system, is associated with amimia, or reduction of spontaneous facial activity (Buck & Duffy, 1980), including a decreased eye blink rate; but subcortical lesions may also lead to increased facial activity, as in Meige's disease or Brueghel's syndrome, thought to result from a disorder of the basal ganglia.

Cortical lesions also influence expression. Lesions of the supplementary motor area (medial part of the frontal lobe) lead to contralateral facial paresis, with spontaneous emotional expression more affected than voluntary; lesions of the motor cortex (also with contralateral facial hemiparesis) affect voluntary movements but leave intact spontaneous smiling (Monrad-Krohn, 1924). Frontal lobe lesions lead to fewer spontaneous expressions of brow raising, smiling, lip tightening, tongue protrusion, etc. during neuropsychological testing of brain injured subjects than parietal or temporal lesions (Kolb & Milner, 1981), though the role of motor versus cognitive or emotional contributions has not been sorted out. The effects of biofeedback, used therapeutically for this condition (Godoy & Carrobles, 1986), could be tracked using facial expression analysis.

Automation of facial measurements could provide the increased reliability, sensitivity, and precision needed to exploit the relationship between facial signs and neurological damage and lead to new insights and diagnostic methods.

Alteration in facial expression may be seen with medical disorders not normally viewed as neurological or psychiatric, such as asthma (Marx et al., 1986), hypertension (Schachter, 1957), conditions associated with pain (Vaughan & Lanzetta 1980; Prkachin & Mercer, 1989). Evaluation of either the production of or the response to facial expression in all these conditions requires measurement of the facial expression produced or presented.

Very recently, anesthesiologists have suggested that it might be possible to detect consciousness during surgery from facial activity. Although the patient experiences no pain and gross motor activity is paralyzed, patients have reported full recall of what was said and other activity in the room during an operation. Not knowing the patient was mentally alert has caused severe problems for both the patients and the medical teams and this provides yet another potential medical application for on-line monitoring of facial activity.

Many congenital disorders and in utero exposures (to prescribed, accidental, and recreational chemicals, as well as infections such as cytomegalovirus) lead to subtle or profound neurological and developmental dysfunction. A comprehensive battery of developmental markers by which to assess and address children's deficits is crucial, and evaluations dealing with production and reception of facial expression are certain to figure in such testing.

Behavioral disorders often accompany (or may occur independently of) minimal or overt brain damage. In this light, the association between delinquency and

inability to recognize facial affects (McCown et al., 1986, 1988) is noteworthy. Assessment of facial expression may serve not only as a marker for dysfunction, but may also aid in management of children who fail to elaborate or fail to heed normal social signals, as suggested by the reduction of disruptive mealtime behavior by facial screening in a mentally retarded girl (Horton, 1987).

Deaf users of sign language have prominent use of facial expression in communication, signaling both meaning and grammar. For instance certain adverbials, or the presence of a relative clause, are indicated by distinctive facial movements (Bellugi, personal communication). There is a dissociation between linguistic and emotive facial expression in deaf signers with right versus left hemisphere lesions. Scientists investigating the use of facial expression in language of deaf and normal speakers have expressed excitement about the possibility of an automated facial expression coding system.

Unforeseen Benefits

These diverse examples illustrate the potential of a focused research program directed towards computer understanding of facial expression. However, the most important applications and benefits could well be the ones that we have not yet imagined. Important strides in measurement methods can give rise to far reaching changes in substantive investigations, and we anticipate many more creative applications to emerge, once this technology is available, that will enhance our understanding of people through their expressions.

VII. REFERENCES.

Abu-Mostafa, Y. S. & Psaltis, D. (1987). Optical neural computers. *Scientific American, 256(3),* 88-95.

Aggarwal, J.K. & Nandhakumar, N. (1988). On the computation of motion from sequences of images - A review. *Proc. of IEEE, 76(8),* 917-935.

Agin, G.J. & Binford, T.O. (1973). Computer analysis of curved objects. *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI),* 629-640.

Aizawa, K., Harashima, H., & Saito, T. (1989). Model-based analysis synthesis image coding (MBasic) system for a person's face. *Signal Processing Image Communication, 1,* 139-152.

Anderson, D. Z. (1986). Coherent optical eigenstate memory. *Opt. Lett., 11(1),* 56-58.

Andreou, A. et. al., (1991). Current mode subthreshold circuits for analog VLSI neural systems. *IEEE-Neural Networks, 2(2),* 205-213.

Azarbayejani, A., Starner, T., Horowitz, B., & Pentland, A. P. (1992). Visual head tracking for interactive graphics. *IEEE Trans. Pattern Analysis and Machine Vision,* special issue on computer graphics and computer vision, in press.

Barnard, S. T. & Thompson, W. B. (1980). Disparity analysis of images. *IEEE trans. PAMI,* PAMI-2, 4, 333-340.

Barrow, H.G. & Popplestone, R.J. (1971). Relational descriptions in picture processing. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence, 6.* Edinburgh Univ. Press, Edinburgh.

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1985). Selectivity between faces in the responses of as population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Research, 342,* 91-102.

Bichsel, M. & Pentland, A. (1992). *Topological matching for human face recognition.* M.I.T. Media Laboratory Vision and Modeling Group Techical Report No. 186, June.

Bledsoe, W. W. (1966). *The model method in facial recognition.* Panoramic Research Inc., Palo Alto, CA, PRI:15, Aug.

Borod, J. C., St. Clair, J., Koff, E., T Alpert, M. (1990). Perceiver and poser asymmetries in processing facial emotion. *Brain & Cognition, 13(2),*167-177.

Brooke, N. M. & Petajan, E. D. (1986). Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. *Proceedings of the International Conference on Speech Input/Output: Techniques and Applications,* London, pp 104-109.

Brooke, N. M. (1989). Visible speech signals: Investigating their analysis, synthesis, and perception. In M. M. Taylor, F. Neel, & D. G. Bouwhuis (Eds.), *The Structure of Multimodal Dialogue*. Holland: Elsevier Science Publishers.

Brunelli, R. (1990). *Edge projections for facial feature extraction*. Technical Report 9009-12, Instituto per la Ricerca Scientifica e Tecnologica, Trento, Italy.

Brunelli, R. (1991). *Face recognition: Dynamic programming for the detection of face outline.* Technical Report 9104-06, Instituto per la Ricerca Scientifica e Tecnologica, Trento, Italy.

Brunelli, R. & Poggio, T. (1991). HyperBF networks for gender recognition. *Proceedings Image Understanding Workshop 1991*. San Mateo, CA: Morgan Kaufmann.

Buck, R. & Duffy, R. (1980). Nonverbal communication of affect in brain-damaged patients. *Cortex, 16,* 351-362.

Buhmann, J., Lange, J., & von der Malsburg, C. (1989). Distortion invariant object recognition by matching hierarchically labeled graphs. *IJCNN International Conference on Neural Networks*, (Vol. I, pp. 155-159).Washington, DC.

Buhmann, J., Lange, J., v.d.Malsburg, C., Vorbruggen, J. C., & Wurtz, R. P. (1991). Object recognition with Gabor functions in the Dynamic Link Architecture -- Parallel implementation on a transputer network. In B. Kosko (Ed.)*, Neural Networks for Signal Processing* (pp. 121--159). Englewood Cliffs, NJ: Prentice Hall.

Burt, P. (1988a). Algorithms and architectures for smart sensing. *Proceedings of the Image Understanding Workshop*, April. San Mateo, CA: Morgan Kaufmann,.

Burt, P. J. (1988b). Smart sensing within a pyramid vision machine, *Proceedings of the IEEE, 76(8),* 1006-1015.

Cacioppo, J. T. & Dorfman, D. D. (1987). Waveform moment analysis in psychophysiological research. *Psychological Bulletin, 102,* 421-438.

Cacioppo, J. T. & Petty, R. (1981). Electromyographic specificity during covert information processing. *Psychophysiology, 18(2)*, 518-523.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1985). Semantic, evaluative, and self-referent processing: Memory, cognitive effort, and somatovisceral activity. *Psychophysiology, 22,* 371-384.

Cacioppo, J. T., Tassinary, L. G., & Fridlund, A. F. (1990). The skeletomotor system. In J. T. Cacioppo and L. G. Tassinary (Eds.), *Principles of*

psychophysiology: Physical, social, and inferential elements (pp. 325-384). New York: Cambridge University Press.

Camras, L. A. (1977). Facial expressions used by children in a conflict situation. *Child Development, 48,* 1431-35.

Cannon, S. R., Jones, G. W., Campbell, R., & Morgan, N. W. (1986). A computer vision system for identification of individuals. *Proceedings of IECON* (pp. 347-351).

Carey, S. & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science, 195,* 312-313.

Chen, H. H. & Huang, T. S. (1988). A survey of construction and manipulation of octrees. *Computer Vision, Graphics, and Image Processing (CVGIP), 43,* 409-431.

Chernoff, H. (1971). *The use of faces to represent points in N-dimensional space graphically.* Office of Naval Research, December, Project NR-042-993.

Chernoff, H. (1973). The use of faces to represent points in K-dimensional space graphically, *Journal of American Statistical Association,* 361.

Chesney, M. A., Ekman, P., Friesen, W. V., Black, G. W., & Hecker, M. H. L. (1990). Type A behavior pattern: Facial behavior and speech components. *Psychosomatic Medicine, 53,* 307-319.

Choi, C. S., Harashima, H., & Takebe, T. (1990). 3-Dimensional facial model-based description and synthesis of facial expressions (in Japanese), trans. *IEICE of Japan, J73-A(7), July,* pp 1270-1280.

Choi, C. S., Harashima, H., & Takebe T. (1991). Analysis and synthesis of facial expressions in knowledge-based coding of facial image sequences. *Internationl Conference on Acoustics Speech and Signal Processing* (pp. 2737-2740). New York: IEEE.

Churchland, P. S. & Sejnowski, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

Cotter, L. K., Drabik, T. J., Dillon, R. J., & Handschy, M. A. (1990). Ferroelectric-liquid-crystal silicon-integrated-circuit spatial light modulator. *Opt. Lett., 15(5),* 291.

Cottrell, G. W. & Fleming, M. K. (1990). Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference* (pp. 322-325).

Cottrell, G. W. & Metcalfe, J. (1991). EMPATH: Face, gender and emotion recognition using holons. In R. P. Lippman, J. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems 3*. (pp. 564-571). San Mateo, CA: Morgan Kaufmann..

Craw, I., Ellis, H., & Lishman, J. R. (1987). Automatic extraction of face features. *Pattern Recognition Letters, 5,* 183-187.

Cyberware Laboratory Inc (1990). *4020/RGB 3D scanner with color digitizer*. Monterey, CA.

Damasio, A., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology, 32,* 331-341.

Darwin, C. (1872). *The expression of the emotions in man and animals.*New York: Philosophical Library.

Davidson, R. J., Ekman, P., Saron, C., Senulis, J., & Friesen, W.V. (1990) Emotional expression and brain physiology I: Approach/withdrawal and cerebral asymmetry. *Journal of Personality and Social Psychology, 58,* 330-341.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience, 3,* 1-24.

Drabik, T. J. & Handschy, M. A. (1990). Silicon VLSI ferroelectric liquid-crystal technology for micropower optoelectronic computing devioces. *Applied Optics 29(35),* 5220.

Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis*. John Wiley.

Eggert, D. & Bowyer, K. (1989). Computing the orthographic projection aspect graph of solids of revolution. *Proc. IEEE Workshop on Interpretation of 3D Scenes*, (pp. 102-108). Austin, TX.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska Symposium on Motivation* 1971, (Vol. 19, pp. 207-283). Lincoln, NE: University of Nebraska Press.

Ekman, P. (1978). Facial signs: Facts, fantasies, and possibilities. In T. Sebeok (Ed.), Sight, Sound and Sense. Bloomington: Indiana University Press.

Ekman, P. (1979). About brows: Emotional and conversational signals. In J. Aschoff, M. von Carnach, K. Foppa, W. Lepenies, & D. Plog (Eds.), *Human ethology* (pp. 169-202). Cambridge: Cambridge University Press.

Ekman, P. (1982). Methods for measuring facial action. In K.R. Scherer and P. Ekman (Eds.), *Handbook of methods in Nonverbal Behavior Research* (pp 45-90). Cambridge: Cambridge University Press.

Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 319-343). Hillsdale, N.J.: Lawrence Erlbaum.

Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. In H. Wagner & A. Manstead (Eds.), *Handbook of social psychophysiology* (pp. 143-164). Chichester: Wiley.

Ekman, P. (1992a). Facial expression of emotion: New findings, new questions. *Psychological Science, 3,* 34-38.

Ekman, P. (1992b). An argument for basic emotions. *Cognition and Emotion, 6,* 169-200.

Ekman, P. & Davidson, R. J. (1992). Voluntary smiling changes regional brain activity. Ms. under review.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). Duchenne's smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology, 58,* 342-353.

Ekman, P. & Fridlund, A. J. (1987). Assessment of facial behavior in affective disorders. In J. D. Maser (Ed.), *Depression and Expressive Behavior* (pp. 37-56). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ekman, P. & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1,* 49- 98.

Ekman, P. & Friesen, W. V. (1975). *Unmasking the face. A guide to recognizing emotions from facial clues.* Englewood Cliffs, New Jersey: Prentice-Hall.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement.* Palo Alto, Calif.: Consulting Psychologists Press.

Ekman, P. & Friesen, W. V. (1986). A new pan cultural facial expression of emotion. *Motivation and Emotion, 10(2),* 1986.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings.* New York: Pergamon Press.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology, 54,* 414-420.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221,* 1208-1210.

Ekman, P. & O'Sullivan, M. (1988). The role of context in interpreting facial expression: Comment of Russell and Fehr (1987). *Journal of Experimental Psychology, 117*, 86-88.

Ekman, P., O'Sullivan, M., & Matsumoto, D. (1991a). Confusions about content in the judgment of facial expression: A reply to Contempt and the Relativity Thesis. *Motivation and Emotion, 15,* 169-176.

Ekman, P., O'Sullivan, M., & Matsumoto, D. (1991b). Contradictions in the study of contempt: What's it all about? Reply to Russell. *Motivation and Emotion, 15,* 293-296.

Ekman, P. & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology, 20,* 527-554.

Ellgring, H. (1989). *Nonverbal communications in depression.* Cambridge: University Press.

Farhat, N., Psaltis, D., Prata, A., & Paek, E. (1985). Optical implementation of the Hopfield Model. *Appl. Opt., 24(10),* 1469-1475.

Finn, K. (1986). *An investigation of visible lip information to be used in automatic speech recognition.* PhD Dissertation, Georgetown University.

Fischler, M.A. & Elschlager, R.A. (1973). The representation and matching of pictorial structures. *IEEE Trans. on Computers, C-22, 1,* 67-92.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11,* 796-804.

Frenkel, K.A. (1994). The human genome project and informatics, *Comm. of the ACM, 34(11),* Nov.

Fridlund, A.J. (1991). Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology, 32,* 3-100.

Fried, L. A. (1976). *Anatomy of the head, neck, face, and jaws.*Philadelphia: Lea and Febiger.

Friedman J.H. & Stuetzle, W. (1981). Projection pursuit regressio. *Journal of the American Statistics Association, 76(376*), 817-823.

Friedman, S. M. (1970). *Visual anatomy: Volume one, head and neck*. New York: Harper and Row.

Friesen, W.V. & Ekman, P. (1987). *Dictionary - Interpretation of FACS Scoring.* Unpublished manuscript.

Garcia, O. N., Goldschen, A. J., & Petajan, E. D. (1992). *Feature extraction for optical automatic speech recognition or automatic lipreading.* Technical Report GWU-IIST-9232, Department of Electrical Engineering and Computer Science, George Washington University, Washington, DC.

Gillenson, M.L. (1974). *The interactive generation of facial images on a CRT using a heuristic strategy*. Ohio State University, Computer Graphics Research Group, The Ohio State University, Research Center, 1314 Kinnear Road, Columbus, Ohio 434210.

Godoy, J. F. & Carrobles, J. A. (1986). Biofeedback and facial paralysis: An experimental elaboration of a rehabilitation program. *Clinical Biofeedback & Health: An International Journal, 9(2),* 124-138.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (in press). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*.

Goldstein, A. J., Harmon, L. D., & Lesk, A. B. (1971). Identification of human faces, *Proceedings of IEEE, 59,* 748.

Golomb, B.A., Lawrence, D.T., & Sejnowski, T.J. (1991). SEXNET: A neural network identifies sex from human faces. In D.S. Touretzky & R. Lippman (Eds.), *Advances in Neural Information Processing Systems, 3,* San Mateo, CA: Morgan Kaufmann.

Govindaraju, V. (1992). *A computational model for face location*. Ph.D. Dissertation, The State University of New York at Buffalo.

Gray, M., Lawrence, D., Golomb, B., & Sejnowski, T. (1993). *Perceptrons reveal the face of sex.* Institute for Neural Computation Technical Report, University of California, San Diego.

Grimson, W. E. L. (1983). An implementation of a computational theory of visual surface interpolation. *Computer Vision, Graphics, and Image Processing (CVGIP) 22 (1),* 39--69.

Gross, C. G., Rocha-Miranda, C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Neurophysiology, 35,* 96-111.

Hager, J. C. (1985). A comparison of units for visually measuring facial action. *Behavior research methods, instruments and computers, 17,*450-468.

Hager, J. C., & Ekman, P. (1985). The asymmetry of facial actions is inconsistent with models of hemispheric specialization. *Psychophysiology, 22(3),* 307-318.

Hall, J. A. Gender effects in decoding nonverbal cues. (1978). *Psychological Bulletin, 85,* 845-857.

Hallinan, P. W. (1991). Recognizing human eyes. *SPIE Proceedings, V. 1570, Geometric Method in Computer Vision,* 214-226.

Harris, J., Koch, C., & Staats, C. (1990). Analog hardware for detecting discontinuities in early vision. *International Journal of Computer Vision, 4(3),* 211-223.

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., & Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences of the United States of America, 88(5)*, 1621-1625.

Henneman, E. (1980). Organization of the motoneuron pool: The size principle. In V. E. Mountcastle (Ed.), *Medical physiology* (14th ed., Vol. 1, pp. 718-741). St. Louis, Mosby.

Heywood, C. A. & Cowey, A. (1992). The role of the 'face-cell' area in the discrimination and recognition of faces by monkeys. *Philosophical Transactions*

of the Royal Society of London. Series B: Biological Sciences, 335(1273), 31-37; discussion 37-38.

Hill, D.R., Pearce, A., & Wyvill, B. (1988). Animating speech: An automated aproach using speech synthesis by rules. *The Visual Computer, 3,* 277-289.

Horn, B. K. P. (1986). *Robot Vision.* McGraw-Hill.

Horn, B. K. P. & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17,* 185-203.

Horton, S. V. (1987). Reduction of disruptive mealtime behavior by facial screening: A case study of a mentally retarded girl with long-term follow-up. *Behavior Modification, 11(1)*, 53-64.

Huang, T. S. (1987). Motion analysis. In S. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* . John Wiley.

Huang, T. S. & Orchard, M. T. (1992). Man-machine interaction in the 21st century: New paradigms through dynamic scene analysis and synthesis, *Proc. SPIE Conf. on Visual Communications and Image Processing '92, Vol. 1818*(pp. 428-429). Nov. 18-20, Boston, MA.

Huang T. S., Reddy, S. C., & Aizawa, K. (1991). Human facial motion modeling, analysis, and synthesis for video compression. *Proceedings of SPIE, 1605,* 234-241.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London), 160,* 106-154.

Hurwitz, T. A., Wada, J. A., Kosaka, B. D., & Strauss, E. H. (1985). Cerebral organization of affect suggested by temporal lobe seizures. *Neurology, 35(9),* 1335-1337.

Izard, C. E. (1971). *The face of emotion.* New York: Appleton-Century-Crofts.

Izard, C. E. (1977). *Human emotions.* New York: Academic Press.

Izard, C. E. (1979). *The maximally discriminative facial movement coding system (MAX).* Unpublished manuscript. Available from Instructional Resource Center, University of Delaware, Newark, Delaware.

Jaeger, J., Borod, J. C., & Peselow, E. (1986). Facial expression of positive and negative emotions in patients with unipolar depression. *Journal of Affective Disorders, 11(1)*, 43-50.

Johnson, M. H. & Morton, J. (1991). *Biology and cognitive development: The case of face recognition.* Oxford, UK; Cambridge, Mass: Blackwell.

Jordan, M. I. & Rumelhart, D. E. (1992). Forward models - Supervised learning with a distal teacher. *Cognitive Science, 16(3),* 307-354.

Kanade, T. (1973). *Picture processing system by computer complex and recognition of human faces.* Dept. of Information Science, Kyoto University, Nov.

Kanade, T. (1977). *Computer recognition of human faces.* Basel and Stuttgart: Birkhauser Verlag.

Kanade, T. (1981). Recovery of the 3D shape of an object from a single view. *Artificial Intelligence 17,* 409-460.

Kass, M., Witkin, A., & Terzopoulos, D. (1987). Snakes: Active contour models. *Proc. ICCV-87,* June, 259-268.

Kaya, Y. & Kobayashi, K. (1972). A basic study of human face recognition. In A. Watanabe (Ed.), *Frontier of Pattern Recognition* (pp. 265).

Kleiser, J. (1989). A fast, efficient, accurate way to represnt the human face. State of the Art in Facial Animation. *ACM, SIGGRAPH '89 Tutorials, 22,*37-40.

Koenderink, J. J. & van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics 32,* 211-216.

Kohonen, T., Lehtio, P., Oja, E., Kortekangas, A., & Makisara, K. (1977). Demonstration of pattern processing properties of the optimal associative mappings. *Proc Intl. Conf. on Cybernetics and Society,* Wash., D.C.

Kolb, B. & Milner, B. (1981). Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia, 19(4),* 491-503.

Komatsu, K. (1988). Human skin capable of natural shape variation. *The Visual Computer, 3,* 265-271,

Krause, R., Steimer, E., Sanger-Alt, C., & Wagner, G. (1989). Facial expression of schizophrenic patients and their interaction partners. *Psychiatry, 52*, 1-12.

Larrabee, W. (1986). A finite element model of skin deformation. *Laryngoscope, 96,* 399-419.

Lee, H. C. (1986). Method for computing the scene-illuminant chromaticity from specular highlights. *Jour. Optical Society of America A (JOSA A) 3 (10),*1694-1699.

Lee, K. F. (1989). *Automatic speech recognition: The development the SPHINX system.* Boston: Kluwer Academic Publishers.

Leung, M. K. & Huang, T. S. (1992). An integrated approach to 3D motion analysis and object recognition, *IEEE Trans PAMI, 13(10),* 1075-1084.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology, 27,* 363-384.

Lewis, J. P. & Parke, F. I. (1987). Automatic lip-synch and speech synthesis for character animation. *CHI+CG '87* , Toronto, 143-147.

Li, H. Y., Qiao, Y., & Psaltis, D. (no date). An optical network for real time face recognition, *Appl. Opt.* Sejnowski/Pentland

Lisberger, S. G. & Sejnowski, T. J. (1992). *Computational analysis predicts the site of motor learning in the vestibulo-ocular reflex.*Technical Report INC-92.1, UCSD.

Magnenat-Thalmann, N., Primeau, N.E., & Thalmann, D. (1988). Abstract muscle actions procedures for human face animation. *Visual Computer, 3(5)*, 290-297.

Mahowald, M. & Mead, C. (1991). The silicon retina. *Scientific American, 264(5)*, 76-82.

Mandal, M. K. & Palchoudhury, S. (1986). Choice of facial affect and psychopathology: A discriminatory analysis. *Journal of Social Behavior & Personality, 1(2),* 299-302.

Maniloff, E. S. & Johnson, K. M. (1990). Dynamic holographic interconnects using static holograms. *Opt. Eng., 29(3),* 225-229.

Marr, D. (1982). *Vision.* San Francisco: W.H. Freeman.

Marx, D., Zofel, C., Linden, U., Bonner, H. et al. (1986). Expression of emotion in asthmatic children and their mothers. *Journal of Psychosomatic Research, 30(5)*, 609-616.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions, E 74, 10,* 3474-3483.

Mase, K. & Pentland, A. (1990a). Lip reading by optical flow, *IEICE of Japan, J73-D-II, 6,* 796-803.

Mase, K. & Pentland, A. (1990b). Automatic lipreading by computer. *Trans. Inst. Elec. Info. and Comm. Eng., J73-D-II(6),* 796-803.

Mase, K. & Pentland, A. (1991). Automatic lipreading by optical flow analysis. *Systems and Computers in Japan, 22(6),* 67-76.

Mase, K., Watanabe, Y., & Suenaga, Y. (1990). A real time head motion detection system. *Proceedings SPIE, 1260,* 262-269,.

McCown, W., Johnson, J., & Austin, S. (1986). Inability of delinquents to recognize facial affects. First International Conference on the Meaning of the Face (1985, Cardiff, Wales). *Journal of Social Behavior & Personality, 1(4)*, 489-496.

McCown, W. G., Johnson, J. L., & Austin, S. H. (1988). Patterns of facial affect recognition errors in delinquent adolescent males. *Journal of Social Behavior &*

Personality, 3(3), 215-224.

McGuigan, F. J. (1970). Covert oral behavior during the silent performance of language tasks. *Psychological Bulletin, 74,* 309-326.

McHugo, G. J., Lanzetta, J. T., Sullivan, D. G., Masters, R. D., & Englis, B. G. (1985). Emotional reactions to a political leader's expressive displays. *Journal of Personality and Social Psychology, 49,* 1513-1529.

McKendall, R. & Mintz, M. (1989). *Robust fusion of location information.* Preprint. Dept. of Computer & Info. Sci., Univ. of Pennsylvania..

Mead, C. (1989). *Anolog VLSI and Neural Systems*. New York: Addison-Wesley Publishing Company.

Moffitt, F.H. & Mikhail, E.M. (1980). *Photogrammetry* (3rd ed.). Harper & Row.

Monrad-Krohn, G. H. (1924). On the dissociation of voluntary and emotional innervation in facial paresis of central origin. *Brain, 47,* 22-35.

Montgomery, A. & Jackson, P. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of Acoustical Society of America, 73(6),* 2134-2144.

Nahas, M., Huitric, H., & Sanintourens, M. (1988). Animation of a B-spline figure. *The Visual Computer, 3,* 272-276.

Nassimbene, E. (1965). U. S. Patent No. 3192321, June 29.

Nishida, S. (1986). Speech recognition enhancement by lip information. *ACM SIGCHI Bulletin, 17(4*), 198-204.

O'Rourke, J. & Badler, N. (1980). Model-based image analysis of human motion using constraint propagation. *IEEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2, 6,* 522-536.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Bartlett, J. C. (1991). Classifying faces by race and sex using an autoassociative memory trained for recognition. In *Proceedings of the Thirteenth Annual Cognitive Science Society*, August 1991, Chicago, IL. pp. 847-851. Hillsdale: Lawrence Erlbaum.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America, A10,* 405-411.

O'Toole, A. J., Millward, R. B., Richard, B., & Anderson, J. A. (1988). A physical system approach to recognition memory for spatially transformed faces. *Neural Networks, 2*, 179-199.

Ohmura, K., Tomono, A., &. Kobayashi, Y. (1988). Method of detecting face direction using image processing for human interface. *Proceedings of SPIE, 1001,* 625-632.

Oka, M. Tsutsui, K., Ohba, A., Kurauchi, Y., & Tago, T. (1987). Real-time manipulation of texture-mapped surfaces. *Computer Graphics, 21(4),* 181-188.

Oster, H., Hegley, D., & Nagel, (in press). Adult judgment and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. *Developmental Psychology*.

Owechko, Y., Dunning, G. J., Marom, E., & Soffer, B. H. (1987). Holographic associative memory with nonlinearities in the correlation domain. *Appl. Opt., 26(10)*, 1900-1910.

Paek, E. G. & Jung, E. C. (1991). Simplified holographic associative memory using enhanced nonlinear processing with a thermalplastic plate. *Opt. Lett., 16(13),* 1034-1036.

Parke, F. I. (1972a). *Computer generated animation of faces.* University of Utah, Salt Lake City, June, UTEC-CSc-72-120.

Parke, F. I. (1972b). Computer generated animation of faces. *ACM Nat'l Conference, 1,* 451-457.

Parke, F. I. (1974). *A parameteric model for human faces.* University of Utah, UTEC-CSc-75-047, Salt Lake City, Utah, December.

Parke, F. I. (1975). A model of the face that allows speech synchronized speech. *Journal of Computers and Graphics, 1,* 1-4.

Parke, F. I. (1982). Parameterized models for facial animation. *IEEE Computer Graphics and Applications, 2(9),* 61-68.

Patel, M. & Willis, P. J. (1991). FACES: Facial animation, construction and editing system. In F. H. Post and W. Barth (Eds.), *EUROGRAPHICS '91* (pp. 33-45). Amsterdam: North Holland.

Pearce, A., Wyvill, B., Wyvill, G., & Hill, D. (1986). Speech and expression: A computer solution to face animation. In M. Wein and E. M. Kidd (Eds.), *Graphics Interface '86* (pp. 136-140). Ontario: Canadian Man-Computer Communications Society.

Pearlmutter, B. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation, 1,* 263-269.

Pelachaud, C. (1991). *Communication and coarticulation in facial animation.* University of Pennsylvania, Department of Computer and Information Science, October.

Pentland, A. (1992). personal communication.

Pentland, A., Etcoff, N., Starner, T. (1992). *Expression recognition using eigenfeatures.* M.I.T. Media Laboratory Vision and Modeling Group Techical Report No. 194, August.

Pentland, A. & Horowitz, B. (1991). Recovery of non-rigid motion. *IEEE Trans. Pattern Analysis and Machine Vision*, *13(7),* 730-742.

Pentland, A. & Mase, K. (1989). *Lip reading: Automatic visual recognition of spoken words.* MIT Media Lab Vision Science Technical Report 117, January 15.

Pentland, A. & Sclaroff, S. (1991). Closed-form solutions for physically-based modeling and recognition *IEEE Trans. Pattern Analysis and Machine Vision,13*(7), 715-730.

Perkell, J. S. (1986). Coarticulation strategies: Preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication, 5(1)* , 47-68.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47,*329-342.

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1984). Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology, 3*, 197-208.

Perry, J. L. & Carney, J. M. (1990). Human face recognition using a multilayer perceptron. *Proceedings of International Joint Conference on Neural Networks,* Washington D.C. Volume 2, pp. 413-416.

Petajan, E. D. (1984). *Automatic lipreading to enhance speech recognition.* PhD Dissertation, University of Illinois at Urbana-Champaign.

Petajan, E. D., Bischoff, B., Bodoff, D., & Brooke, N. M. (1988). An improved automatic lipreading system to enhance speech recognition. *CHI 88,* 19-25.

Pieper, S. D. (1989). *More than skin deep: Physical modeling of facial tissue.* Massachusetts Institute of Technology, 1989, Media Arts and Sciences, MIT.

Pieper, S. D. (1991). *CAPS: Computer-aided plastic surgery.*Massachusetts Institute of Technology, Media Arts and Sciences, MIT, September.

Pigarev, I. N., Rizzolatti, G., & Scandolara, C. (1979). Neurones responding to visual stimuli in the frontal lobe of macaque monkeys. *Neuroscience Letters, 12,* 207-212.

Platt, S. M. (1980). *A System for Computer Simulation of the Human Face.*,The Moore School, 1980, Pennsylvania.

Platt, S. M. (1985). *A structural model of the human face.* The Moore School, Pennsylvania.

Platt, S. M. & Badler, N. I. (1981). Animating facial expressions. *Computer Graphics, 15*(3), 245-252.

Poeck, K. (1969). Pathophysiology of emotional disorders associated with brain damage. In P. J. Vinken & G. W. Bruyn (Eds.), *Handbook of Clinical Neurology V.3* Amsterdam: North Holland.

Poggio, T. (1990). A theory of how the brain might work. In *Cold Spring Harbor Symposia on Quantitative Biology* (pp. 899-910). Cold Spring Harbor Laboratory Press..

Ponce, J. & Kriegman, D. J. (1989). On Recognizing and positioning curved 3D objects from image contours. *Proc. IEEE Workshop on Interpretation of 3D Scenes,* Austin, TX.

Prkachin, K. M. & Mercer, S. R. (1989). Pain expression in patients with shoulder pathology: validity properties and relationship to sickness impact. *Pain, 39,* 257-265.

Psaltis, D., Brady, D., Gu, X.-G., & Lin, S. (1990). Holography in artificial neural networks. *Nature, 343(6256),* 325-330.

Psaltis, D., Brady, D., & Wagner, K. (1988). Adaptive optical networks using photorefractive crystals. *Appl. Opt., 27(9),* 1752-1759.

Psaltis, D. & Farhat, N. (1985). Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. *Opt. Lett., 10(2),* 98-100.

Reeves, A. G. & Plum, F. (1969). Hyperphagia, rage, and dementia accompanying a ventromedial hypothalamic neoplasm. *Archives of Neurology, 20(6)*, 616-624.

Requicha, A.A.G. (1980). Representations of rigid solids. *ACM Computing Surveys, 12,* 437-464.

Rhiner, M. & Stucki, P. Database Requirements for Multimedia Applications. In L. Kjelldahl (Ed.), *Multimedia: Systems, Interaction and Applications*, Springer, 1992.

Rinn, W. E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions.*Psychological Bulletin, 95,* 52-77.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology, 3,* 209-222.

Rolls, E. T., Baylis, G. C., & Hasselmo, M. E. (1987). The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. *Vision Research, 27(3),*311-326.

Rolls, E. T., Baylis, G. C., Hasselmo, M. E., & Nalway, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 76(1),* 153-164.

Rosenfeld, H. M. (1982). Measurement of body motion and orientation. In K.R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research.* (pp. 199-286). Cambridge: Cambridge University Press.

Rumelhart, D. E., Hinton, G., & Williams, R. J. (1986). Learning internal representation by error propagation. In D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing, Explorations in the microstructure of cognition* (pp. 318-362). Cambridge, Mass.: MIT Press.

Russell, J. A. (1991a). The contempt expression and the relativity thesis. *Motivation and Emotion, 15,* 149-168.

Russell, J. A. (1991b). Rejoinder to Ekman, O'Sullivan and Matsumoto. *Motivation and Emotion, 15,* 177-184.

Russell, J. (1991c). Negative results on a reported facial expression of contempt. *Motivation and Emotion, 15,* 281-291.

Russell, J. A. & Fehr, B. (1987). Relativity in the perception of emotion in facial expressions. *Journal of Experimental Psychology, 116,*233-237

Sakai, T., Nagao, M., & Fujibayashi, S. (1969). Line extraction and pattern detection in a photograph, *Pattern Recognition, 1,* 233-248.

Sakai, T., Nagao, M., & Kanade, T. (1972). Computer analysis and classification of photographs of human faces. *First USA-JAPAN Computer Conference,* session 2-7.

Satoh, Y., Miyake, Y., Yaguchi, H., & Shinohara, S. (1990). Facial pattern detection and color correction from negative color film, *Journal of Imaging Technology, 16(2*), 80-84.

Schachter, J. (1957). Pain, fear and anger in hypertensives and normotensives: a psychophysiological study. *Psychosomatic Medicine, 19,* 17-29.

Sejnowski, T. J. & Churchland, P. S. (1992). Silicon brains. *Byte, 17(*10), 137-146.

Sejnowski, T. J. & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145-168.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain, 115(Pt 1)* , 15-36.

Sethi, I. K. & Jain, R. (1987). Finding trajectories of feature points in a monocular image sequence. *IEEE trans. PAMI,* PAMI-9(1), 56-73.

Smith, S. (1989). Computer lip reading to augment automatic speech recognition. *Speech Techology,* 175-181.

Steimer-Krause, E., Krause, R., & Wagner, G. (1990). Interaction regulations used by schizophrenics and psychosomatic patients. Studies on facial behavior in dyadic interactions. Psychiatry, 53, 209-228.

Stern, J. A., & Dunham, D. N. (1990). The ocular system. In J. T. Cacioppo and L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 513-553). New York: Cambridge University Press.

Stork, D. G., Wolff, G., & Levine, E. (1992). Neural network lipreading system for improved speech recognition. *Proceedings of the 1992 International Joint Conference on Neural Networks*, Baltimore, MD.

Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustic Society of America, 26,* 212-215.

Tanner, J. E. & Mead, C. A. (1984). A correlating optical motion detector. In Penfield (Ed.), *Proceedings of Conference on Advanced Research in VLSI*, January, MIT, Cambridge, MA.

Tassinary, L. G., Cacioppo, J. T., & Geen, T. R. (1989). A psychometric study of surface electrode placements for facial electromyographic recording: I. The brow and cheek muscle regions. *Psychophysiology, 26,* 1-16.

Terzopoulos, D. & Waters, K. (1990a). Analysis of facial images using physical and anatomical models. *Proceedings of the International Conference on Computer Vision,* 1990, 727-732.

Terzopoulos, D. & Waters, K. (1990b). Physically-based facial modeling, analysis, and animation. *Journal of Visualization and Computer Animation, 1(4),* 73-80.

Tolbruck, T. (1992). *Analog VLSI visual transduction and motion processing*. Caltech Ph.D. thesis.

Tranel, D., Damasio, A. R., Damasio, H. (1988). Intact recognition of facial expression, gender and age inpatients with impaired recognition of face identity. *Neurology, 38,* 690-696.

Turk, M. A. (1991). *Interactive time vision: Face recognition as a visual behavior.* Ph.D. Thesis, MIT.

Turk, M. A. & Pentland, A.P. (1989). Face processing: Models for recognition. *SPIE, Intelligent Robots and Computer Vision VIII, 192.*

Turk, M. A. & Pentland, A. P. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3(1),* 71-86.

Van Dongen, H. R., Arts, W. F., & Yousef-Bak, E. (1987). Acquired dysarthria in childhood: An analysis of dysarthric features in relation to neurologic deficits. *Neurology, 37(2)* , 296-299.

Van Gelder, R. S. & Van Gelder, L. (1990). Facial expression and speech:

Neuroanatomical considerations. Special Issue: Facial asymmetry: Expression and paralysis. *International Journal of Psychology, 25(2),* 141-155.

Vannier, M. W., Pilgram, T., Bhatia, G., & Brunsden, B. (1991). Facial surface scanner. *IEEE Computer Graphics and Applications, 11(6),*72-80.

Vaughn, K. B., & Lanzetta, J. T. (1980). Vicarious instigation and conditioning of facial expressive and autonomic responses to a model's expressive display of pain. *Journal of Personality and Social Psychology, 38,* 909-923.

Viennet, E. & Fogelman-Soulie, F. (1992). Multiresolution scene segmentation using MLPs. *Proceedings of International Joint Conference on Neural Networks*, V. III (pp. 55-59). Baltimore.

Wagner, K. & Psaltis, D. (1987). Multilayer optical learning networks, *Appl. Opt, 26(23),* 5061-5076.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37,* 328-339.

Waite, C. T. (1989). *The Facial Action Control Editor, Face: A Parametric Facial Expression Editor for Computer Generated Animation.* Massachusetts Institute of Technology, Media Arts and Sciences, Cambridge, Febuary.

Waite, J. B. & Welsh, W. J. (1990). Head boundary location using snakes, *British Telecom Technol. J.* 8(3), 127-136.

Wang, S.-G. & George, N. (1991). Facial recognition using image and transform representations. In *Electronic Imaging Final Briefing Report*, U.S. Army Research Office, P-24749-PH, P-24626-PH-UIR, The Institute of Optics, University of Rochester, New York.

Watanabe, Y. & Suenaga, Y. (1992). A trigonal prism-based method for hair image generation. *IEEE Computer Graphics and Applications,* January, 47-53.

Waters, K. (1986). Expressive three-dimensional facial animation. *Computer Animation (CG86)* , October, 49-56.

Waters, K. (1987). A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH'87), 21(4),* July, 17-24.

Waters, K. (1988). *The computer synthesis of expressive three-dimensional facial character animation.* Middlesex Polytechnic, Faculty of Art and Design, Cat Hill Barnet Herts, EN4 8HT, June.

Waters, K. & Terzopoulos, D. (1990). A physical model of facial tissue and muscle articulation. *Proceedings of the First Conference on Visualization in Biomedical Computing,* May, 77-82.

Waters, K. & Terzopoulos, D. (1991). Modeling and animating faces using scanned data. *Journal of Visualization and Animation, 2(4),* 123-128.

Waters, K. & Terzopoulos, D. (1992). The computer synthesis of expressive faces. *Phil. Trans. R. Soc. Lond., 355(1273)*, 87-93.

Welsh, W. J., Simons, A. D., Hutchinson, R. A., Searby, S. (no date). Synthetic face generation for enhancing a user interface, British Telecom Research Laboratories, Martlesham Heath, Ipswich IP5 7RE, UK.

Will, P. M. & Pennington, K. S. (1971). Grid Coding: A preprocessing technique for robot and machine vision. *Artificial Intelligence, 2,* 319-329.

Williams, L. (1990). Performace driven facial animation. *Computer Graphics, 24(4),* 235-242.

Witkin, A. P. (1981). Recovering surface shape and orientation from texture. *Artificial Intelligence 17,* 17-45.

Wolff, L. B. (1988). Shape from photometric flow fields. *Proc. SPIE, Optics, Illumination, and Image Sensing for Machine Vision, III,* (pp. 206-213) Cambridge, MA.

Wong, K. H., Law, H. H. M., & Tsang, P. W. M. (1989). A system for recognizing human faces, *Proceedings of ICASSP,* May, pp. 1638-1641.

Wyvill, B. (1989). Expression control using synthetic speech. *State of the Art in Facial Animation, SIGGRAPH '89 Tutorials, ACM, 22,* 163-175.

Yamamoto, M. & Koshikawa, K. (1991). Human motion analysis based on a robot arm model. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 664-665). June, Maui, Hawaii.

Yamana, T. & Suenaga, Y. (1987). *A method of hair representation using anisotropic reflection.* IECEJ Technical Report PRU87-3, May, 15-20, (in Japanese).

Yamato, J., Ohya, J., & Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Champaign, Illinois, 379-385.

Yeh, P., Chiou, A.E.T., & Hong, J. (1988). Optical interconnection using photorefractive dynamic holograms. *Appl. Opt., 27(11),* 2093-2096.

Young, M. P. & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science, 56,* 1327-1331.

Yuhas, B., Goldstein, M. Jr., & Sejnowski, T. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine,* November, 65-71,

Yuille, A. L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience, 3(1),* 59-70.

Yuille, A. L., Cohen, D. S., & Hallinan, P. W. (1989). Feature extraction from

faces using deformable templates. *IEEE proc. CVPR*(June), 104-109.

Yuille, A. L. & Hallinan, P. W. (1992). Deformable Templates. In A. Blake & A.L. Yuille (Eds.), *Active Vision.* Cambridge, Mass: MIT Press.

Zajonc, R. B. (1984). The interaction of affect and cognition. In K. R. Scherer and P. Ekman (Eds.), *Approaches to Emotion* (pp. 239-246). Hillsdale, N.J.: Lawrence Erlbaum Associates.