1.0 INTRODUCTION

Human face is a rich and powerful source of communicative information about human behavior. Facial expression provides sensitive cues about emotional response and plays a major role in human interaction and nonverbal communication. It can complement verbal communication, or can convey complete thoughts by itself. It displays emotion [(35)*], regulates social behavior [(23)], signals communicative intent [(38)], is computationally related to speech production [(70)], and may reveal brain function and pathology [(81)]. Thus, to make use of the information afforded by facial expressions, automated reliable, valid, and efficient methods of measurement are critical.

Computer-vision based approaches to facial expression analysis discriminate among a small set of emotions [(12,13,36,68,82,103)]. This focus follows from the work of Darwin [(32)] and more recently Ekman [(35)], who proposed "six basic emotions" (*i.e.*, joy, surprise, anger, sadness, fear, and disgust), each of which has a prototypic facial expression involving changes in facial features in multiple regions of the face. These basic expressions, however, occur relatively infrequently in everyday life and emotion expression is far more varied. Facial action (detailed facial motion) more often is communicated by subtle changes in one or several discrete features, such as tightening the lips which may communicate anger. In reality, humans are capable of producing thousands of expressions that vary in complexity, intensity, and meaning.

There is a standard anatomically based Facial Action Coding System (FACS) [(34)] developed by psychologists for use in coding facial expressions. With FACS, observers can manually code all possible discrete movements of the face, which are referred to as action units (AUs). AUs individually or in combination can represent all visibly discriminable expressions. Considering the complication of the movements involved and

* Parenthetical references placed superior to the line of text refer to the bibliography.

discrimination of the subtle changes, there is a need to develop an automated system for efficient and quantitative measurement of facial expressions based on FACS, which will make the standardized facial expression measurements more accessible for research in various fields including cognitive and behavior science, psychology, biomedical engineering, teleconferencing and human-computer interface or interaction (HCI).

At the present time, most active communication between human and computer is still in one direction: from human to computer, even though computers may hear human speeches through the use of special audio and speech recognition equipments. Allowing computers to understand human operators through vision will bridge the gap of active communication from the direction of computer to human, which will make computers more active, smart, and friendly. Human face is the richest source of nonverbal communication and the most accessible interface displaying human emotion. To automatically analyze and recognize facial expressions using computers will revolutionize fields which rely on human-computer interaction so that computers will be able to understand whether users feel excited or bored, agrees or disagrees. It will be a great challenge and of practical significance to develop a computer vision system which can automatically recognize a variety of facial expressions and estimate expression intensity.

1.1 Related Works

An increasing number of researchers in computer vision have developed various techniques in providing capabilities for automatic facial expression recognition. We will briefly review the strengths and weaknesses of some major paradigms.

Three-dimensional geometric wireframe (or mesh) face models have been used by Aizawa, Harashima and Saito [1]; Choi, Harashima and Takebe [21]; Essa and Pentland [36]; and Terzopoulos and Waters [91] for facial expression analysis, synthesis and recognition. Essa and Pentland [36] developed two methods to recognize expressions. The first method is to recognize 5 expressions: smile, surprise, raised brow, anger, and disgust by scoring

the dot-product similarity based on 36 peak muscle actuations in comparison to the standard training expression templates but the temporal affect is ignored. The overall recognition rate was 97.8% on 6 subjects with 23 and 48 image sequences for training and testing, respectively. The second method uses the temporal-template matching for two-dimensional gray value images. The time warping is an important consideration which improves the recognition accuracy since the temporal-template matching measures the correlation between testing and standard template image sequences.

In contrast to the use of the complex three-dimensional geometric models, Himer, Schneider, Kost, and Heimann [46]; and Kaiser and Wherle [48] proposed a method for automated detection of facial actions by tracking the positions of attached dots on the face as appeared in an image sequence. Since the shape of dots will deform due to muscle movement during facial expression, it is difficult to locate accurately the corresponding central positions for the deformed dots, and thus affects the tracking accuracy.

Optical flow in two dimensions has been used to track motion and classify basic emotion expression (Black, Yacoob, Jepson and Fleet [12,13]; Mase and Pentland [67,68]; Rosenblum, Yacoob and Davis [82]; and Yacoob and Davis [103]). In work by Mase [68], motions of facial muscles were computed rather than those of facial features. Muscle regions were manually selected by referring to major feature points in the face. Optical flow was computed to extract 12 of the 44 facial muscle movements, which in combination with feature positions were interpreted as appropriate AUs. Mase's approach relies heavily on accurate tracking of the manually selected muscle regions; flow directions within each individual region is averaged to represent the flow direction of that region. However, when the selected area corresponds to a smooth, featureless surface in the face, the optical flow estimation will be unreliable, leading to tracking error. Some selected muscle regions may be difficult to locate manually since they are small and highly mobile. In essence, Mase built a model that is appropriate for synthesizing facial expressions but remains uncertain in analyzing facial expressions. He computed mean and covariance of the optical flow in each local region, and then, based on the highest ratio of between-class

to within-class variability to classify various expressions; the k-nearest-neighbor rule was applied for recognition. His experiments indicated an accuracy of approximately 86% in recognizing five expressions (happiness, anger, surprise, disgust, and unknown) on 1 subject with 20 and 30 training and testing image sequences, respectively.

The work of Yacoob and Davis [103]; and Rosenblum, Yacoob and Davis's [82] are related closely to Mase's in that they used optical flow to track the motion of the surface regions of facial features: brows, eyes, nose and mouth, but not that of the underlying muscle groups. In each facial feature region, the flow magnitude was thresholded to reduce the effect of small computed motions which may be either produced from textureless parts or affected by illumination. The overall flow direction of each region is to conform with the plurality in the neighborhood. The direction of any flow in this region is quantized to one of eight main directions to give a mid-level representation (to match with the dictionary or lookup table of the motion direction for each region of the basic facial action) so as to permit the high-level classification of facial expressions. Yacoob and Davis [103] used this mid-level representation to classify the six basic facial expressions as well as eye blinking. The recognition rate was 88% (except eye blinking for which it was 65%) among 32 subjects with 46 image sequences. Rosenblum, Yacoob and Davis [82] extended Yacoob and Davis's [103] work, based on the similar mid-level representation to recognition of facial expressions of smiling and surprise, using an artificial neural networks with radial basis function (RBF). The recognition rate achieved was 88% for 32 subjects.

Black and Yacoob [12] used a local parameterized model of the image motion to separate and recognize the non-rigid facial expression from the rigid head motion. Their high-level recognition approach was similar to that of Yacoob and Davis's technique [103], which is based on the mid-level index of the motion direction of each facial feature region (brows, eyes and mouth). The mid-level representation was predicted, however, by taking the difference of the motion parameter estimation and a threshold value. Thresholding motion parameters would filter out some subtle motion. Furthermore, different threshold

were used for different motion parameters in the experiment [12]: some were between 0.5 ~ -0.5, and others were between 0.00005 ~ -0.00005. This thresholding method for motion parameters, in effect, reduced reliability and accuracy of the recognition. In their studies of recognition of six basic facial expressions, the average recognition rate was 92% in 40 subjects with 70 image sequences.

Principal component analysis (PCA) has been used previously in gray-value base for recognition expressions of on the forehead and brows (Bartlett, Viola, Sejnowski, Golomb, Larsen, Hager, and Ekman [4]), for face recognition (Kirby and Sirovich [54]; and Turk and Pentland [93]), and for object recognition with varying poses (rigid motion) and illumination (Murase and Nayar [73]). It has also been used in optical flow base for recognition of smile and mouth motion (Black, Yacoob, Jepson, and Fleet [13]), and of lip-reading (Mase and Pentland [67]). Black, et. al. [13] assigned thresholds for motion parameters of linear combination in PCA for their classification paradigm. Mase, et. al. [67] considered the averaged flow direction at each of four rectangular feature regions around the mouth for lip-reading recognition. Both of these constraints introduced some degree of insensitivity to the extracted motion information and thus limited the recognition ability and accuracy

Mase and Pentland's lip-reading approach [67] uses a template matching that minimizes the sum of squared differences (SSD) between the projected flow curve of the testing word and that of the word templates in the two-dimensional eigenspace. The work of Black, Yacoob, Jepson, and Fleet on smile and mouth motion recognition [13] uses a similar approach by comparing similarities of the parameters of linear combination in PCA between the training and testing image sequences. In both cases, time warping is an essential preprocessing for comparison purpose. This becomes impractical when the lengths of image sequences are arbitrary (say, from 9 to 47 frames) and the projected flow curves are in a higher dimensional eigenspace.

Kobayashi and Hara [55,56,57] used three sets of artificial neural networks to recognize six basic facial expressions, mixed facial expressions (combinations of 2 or 3 basic

components), and the intensity of each facial expression, respectively. Inputs to these neural networks are the movements of sixty facial characteristic points which are manually selected. The recognition rate for six basic facial expressions was 88.7% from 15 subjects, and 70% for recognizing mixed facial expressions from 10 subjects.

Other studies in Japan [33,44,80,85,89,94] have used approaches similar to that of Kobayashi and Hara based on the displacement of manually selected facial characteristic points. Ding, Shimamura, Kobayashi, and Nakamura [33] used three sets of artificial neural networks to recognize brows, eyes and mouth expressions. They assumed symmetrical facial expressions, so they performed recognition only on the left half of the face. Others used fuzzy logic (Hashiyama, Furuhashi, Uchikawa and Kato [44], Ralescu and Hartani [80], and Ushida, Takagi and Yamaguchi [94]) or chaos (Sato and Yamaguchi [85]) combined with artificial neural networks to recognize six basic facial expressions.

Bartlett, Viola, Sejnowski, Golomb, Larsen, Hager, and Ekman [4] used three methods (PCA of difference images, optical flow with correlation coefficients, and high gradient component, *i.e.*, wrinkle, detection) to extract information on upper facial expressions (six upper face FACS AUs: AU1, 2, 4, 5, 6 and 7), and employed artificial neural networks for recognition. To deal with the time warping problem, they proposed to manually pick up six frames from each image sequence to form a new sequence for further processing: neutral expression for the first frame, low magnitude expressions for the second frame, medium magnitude expressions for the third and forth frames, and high magnitude expressions for the last two frames. Considering the relative geometric correspondence of face images, they took care of only rotation and horizontal scaling based on the location of both eyes, which was insufficient for aligning face images accurately because the vertical scaling was missing and the sizes of faces could be very different among subjects. The information they used for the PCA is the differences in images obtained by subtracting the gray values of the neutral expression (first frame) from those of the subsequent images for each image sequence. Such a simple subtraction process is not adequate to take care of the differences among individual faces. For high gradient component detection, they did

not discriminate that some wrinkles may be produced by facial expressions while others may be a permanent characteristic of the individual's face. Also, they proposed to wrinkles along several lines where some subjects may and other subjects may not appear wrinkled with the same expression. The best recognition rate was 91% from 20 expert subjects with 80 image sequences and 400 images.

Other methods of recognition have been applied to face or facial expression analysis. Beymer [9] proposed a method to normalize face images across different subjects, he analyzed and synthesized face images by interleaving shape and texture computations using optical flow and PCA in gray-value base [10]. Bregler and Konig [15] employed PCA and Hidden Markov Model (HMM) for speech recognition (eigenlips) and other applications. Kanade [49], one of the pioneers in face identification, used geometrical features of the face, such as the length of facial features, distance between features, and chin shape, to identify a face. Samaria and Young [84] converted each two-dimensional static and mono-shot face image into a concatenated one-dimensional gray-value vector for use in a continuous HMM to identify a face. These are indirectly related to our study but provide valuable references to this research.

1.2 Problem Statement

As reviewed in the previous section, most research in facial expression recognition is limited to six basic expressions and several combinations [12,13,36,68,82,103]. These stylized expressions are classified into emotion categories rather than facial actions. It is insufficient to describe all facial expressions because, in everyday life, six basic expressions occur relatively infrequently. Emotion is often communicated by small changes in one or two discrete features; on the other hand, the same facial expression may be involved in more than one emotion. The presence or absence of one or more facial actions may change its interpretation. For example, as shown in Figure 1, different smile expressions have an action unit AU12 (lip corners pulled obliquely) and emotional simile which may

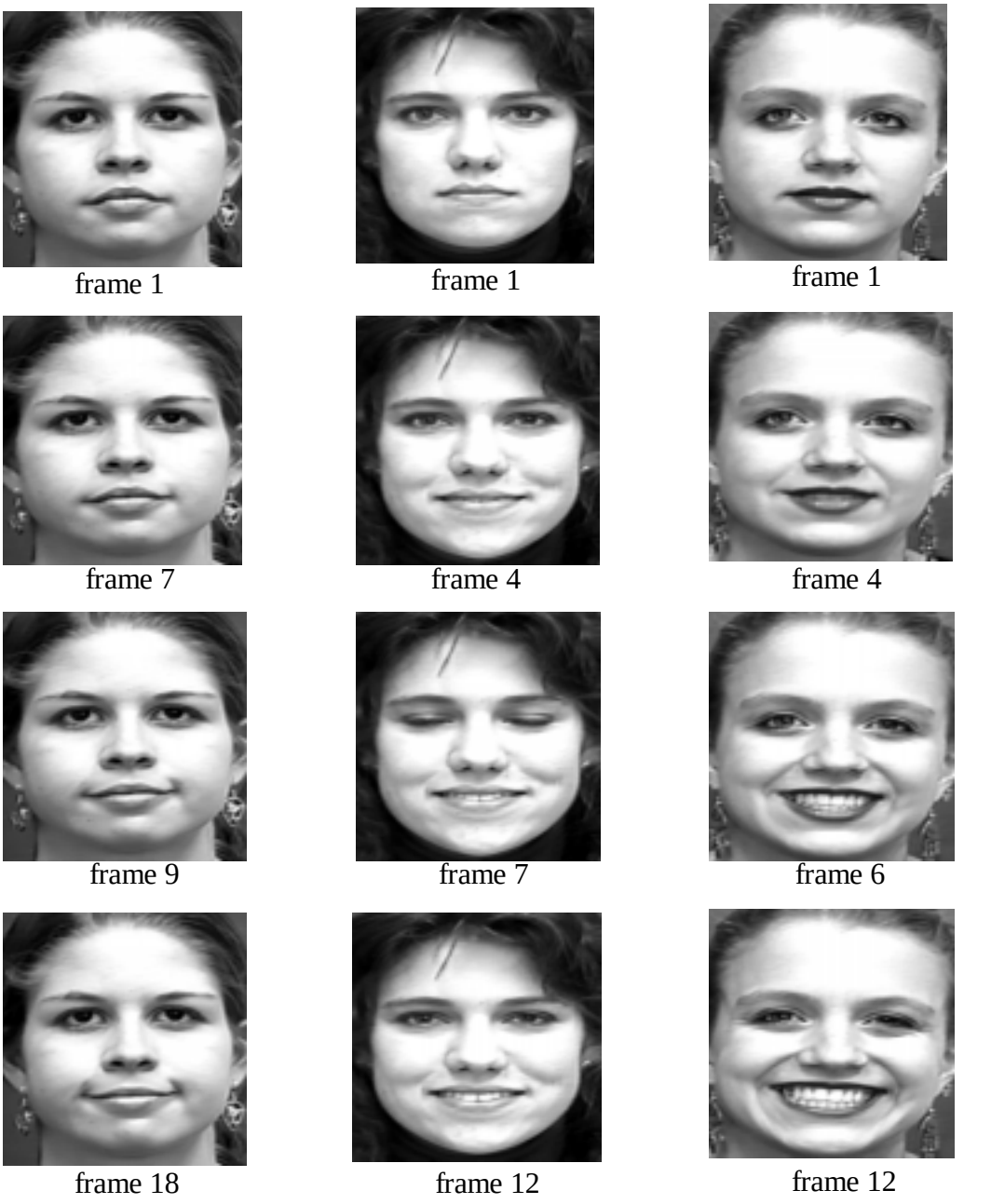

Figure 1 Comparison of different smile expressions with different expression intensities. The presence or absence of one or more facial actions can change their interpretations.

indicate an anxious or concealed emotion, a grin or a genuine (AU6+12+25: lip corners pulled obliquely for AU12, cheek raised for AU6, and subtly exposed for AU25, with or without crow-feet wrinkles near corners of the eyes). The degree of smiling is communicated by the intensity of raising the cheek and lip corners, and having the wrinkles. So it is important to be able not only to recognize basic expressions but also to discriminate subtly different expressions and estimate their expression intensities, which have a similar gross morphology but indicate varied meanings.

The Facial Action Coding System (FACS) [34] is so far the most comprehensive method of coding facial expressions, and provides a guideline for discrimination among closely related expressions. Manually encoding all action units (AUs) for various facial expressions is a laborious process. It takes approximately 100 hours to train a technician to have acceptable levels of coding experiences and up to 10 hours to code one-minute video tape of facial behavior [4,26]. Thus, it is desirable to automate the extraction and coding process, capable of delineating the temporal dynamics and intensity of facial expressions. Feature points are to be tracked in pixel base instead of averaging flow directions in a feature region. Optical flow in a larger region is to be efficiently represented for discriminating expressions. Image sequences ought to be preprocessed to separate the non-rigid motion of facial expression from any rigid head motion as much as possible and to geometrically normalize (align) corresponding face images in a sequence to ensure the assessment of correct motion information.

As the facial motion information are encoded into symbol sequences, it is natural to consider an HMM for automatic recognition of facial expressions where the maximum likelihood decision is assigned to an observable expression symbol sequence. HMM is capable of taking care the problem of variable expression length. HMM topology is referred to a particular network of states and state transitions. The best model is one with as few parameters as possible that can capture the behavior of the training data set. There exists no unified method to determine the optimum topology for an HMM. It will be a

great challenge to develop a strategy to do so for the underlying facial expression recognition problem.

1.3 Objective of the Research

The objective of this dissertation research is to develop a computer vision system, including both facial feature extraction and facial expression recognition based on FACS AUs, that is capable of automatically discriminating among subtly different facial expressions.

For facial feature extraction, we will consider three approaches in parallel. We will apply a pixel-based feature point tracking method, based on the coarse-to-fine pyramid approach, so as to make it sensitive to subtle feature motion as well as to handle large displacements; it will produce facial expression descriptions corresponding to each individual AU or AU combinations. We will also develop a method by using the dense flow to track motion vectors over a large facial region and applying the principle component analysis for data compression yet yielding the entire facial motion information. In addition, we will extract and analyze the motion of high gradient components (furrows) in the spatio-temporal domain to exploit their transient variances associated with facial expression.

Upon extraction of the facial expression information, each motion vector sequence will be vector quantized to a symbol sequence to provide an input to the facial expression classifier. An HMM-based classifier will be designed to deal with varies of facial expressions which are to be recognized in the context of motion sequences of variable length. Furthermore, different methods based on different types of the extracted expression information will be developed for expression intensity estimation which will be useful to segment facial expression sequences, measure the meaning of the expression, and analyze and synthesize facial expression for MPEG-4 applications in teleconferencing.

1.4 Organization of the Dissertation

The dissertation is organized into nine chapters. Chapter 1 gives the motivation of this research and reviews briefly the related works on facial expression recognition systems. The objectives of this dissertation are discussed. Chapter 2 introduces the framework of our computer vision system for facial expression recognition where FACS is used as a basis and where feature tracking is contemplated. Under certain limitations, the rigid head motion is removed from non-rigid facial expressions, and a geometric normalization is prescribed to ensure that optical flows or gray values of face images have the close geometric correspondence.

Chapters 3 through 5 present three methods to extract detail information of facial expressions and to give expression intensity estimation. Chapter 3 describes the pixel-based facial feature point tracking method using the pyramid approach for extracting subtle as well as large movements of facial features in subpixel accuracy. The critical role of the window function in motion estimation is analyzed. The motion is vector quantized into a symbol sequence representing the facial expression. Chapter 4 employs the wavelet-based motion estimation technique for dense flow tracking in order to include information of the entire range of facial motion. Flow-based principal component analysis is presented to compress the high-dimensional dense flows to a low-dimensional weight vector for each frame in a video sequence, which is encoded for recognition processes and expression intensity estimation. Chapter 5 presents a technique for high gradient component (*i.e.*, furrows) analysis. Motion line and edge detectors are designed to extract high gradient components in the spatio-temporal domain and to distinguish furrows from the noise. The high gradient components are encoded to mean and variance vectors as inputs to the recognition process.

Chapters 6 and 7 present the HMM for facial expression recognition. Chapter 6 analyzes the HMM technique and its associated computational issues. Chapter 7 presents a method of determining a special HMM topology for applications to facial expression recognition.

Chapter 8 describes our experimental results. A large database has been tested, subjects ranged in gender, age and ethnicity. We analyzed the performances of the recognition system using three feature tracking methods and demonstrated its high accuracy in comparison to the ground truth by human observation.

Chapter 9 discusses our major contributions and suggestions for further research.