Recognizing linguistic non-manual signs in Sign Language

NGUYEN TAN DAT

(B.Sc. in Information Technology,

University of Natural Sciences,

Vietnam National University - Ho Chi Minh City)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2011

Acknowledgment

A Ph.D. program is a long and difficult trip which I cannot finish without supports and encouragements of many people.

I would like to thank A/P. Surendra Ranganath deeply for his constant guidance and support during this Ph.D. work. His principles for doing research always encourage me to learn more and achieve better in my present and future research.

I would like to express my gratitude to A/P. Ashraf Kassim and Prof. Y.V. Venkatesh for their valuable supports and discussions.

I am grateful to Ms. Judy Ho and other members of Deaf and Hard-of-Hearing Foundation of Singapore for providing me precious knowledge and data of sign language.

My thank also goes to the laboratory technician Mr. Francis Hoon for providing me with all necessary technical supports.

I thank my friends and colleagues for sharing my up and down times: Sylvie, Linh, Chern-Horng, Litt Teen, Loke, Wei Weon, and a lot of others.

Finally, I specially thank Shimiao for her love and supports during these years. My parents, I thank you for your quiet love and sacrifices to make me and this thesis possible.

Contents

Summary

Besides manual (hand) signs, non-manual signs (facial, head, and body behaviors) play an important role in sign language communication used by the deaf. Non-manual signs can be used to convey feelings, linguistic information, etc. In this thesis, we focus on recognizing an important class of non-manual signals in American Sign Language (ASL): grammatical markers which are facial expressions composed of facial feature movements and head motions and are used to convey the structure of a signed sentence. Without satisfactory recognition of grammatical markers, any sign language recognition system cannot fully reconstruct a signed sentence. Six common grammatical markers are considered in this thesis: Assertion, Negation, Rhetorical question, Topic, Wh question, and Yes/no question. These can be identified by combined analysis of facial feature movements and head motions. While there have been attempts in the literature to recognize head movements alone or facial expressions alone, there are few works which consider recognizing facial expressions with concurrent head motion. Indeed, in the facial expression recognition literature, most works assume that the face is frontal with little or no head motion, and most attention has been focused on recognizing the six universal expressions (anger, disgust, fear, happiness, sadness, and surprise). However, in fa-

cial expressions used in sign language, meaning is jointly conveyed through both channels, facial expression (through facial feature movements), and head motion.

In this thesis, we address the problem of recognizing the six grammatical marker expressions in sign language. We propose to track facial features through video, and extract suitable features from them for recognition. We developed a novel tracker which uses spatio-temporal face shape constraints, learned through probabilistic principal component analysis (PPCA), within a recursive framework. The tracker has been developed to yield robust performance in the challenging sign language domain where facial occlusions (by hand), blur due to fast head motion, rapid head pose changes and eye blinks are common. We developed a database of facial video using volunteers from the Deaf and Hard of Hearing Federation of Singapore The videos were acquired while the subjects were signing sentences in ASL.

The performance of the tracker has been evaluated on these videos, as well as on videos randomly picked from the internet, and compared with the Kanade-Lucas-Tomasi (KLT) tracker and some variants of our proposed tracker with excellent results. Next, we considered isolated grammatical marker recognition using an HMM-SVM framework. Several HMMs were used to provide the likelihoods of different types of head motion (using features at rigid facial locations) and facial feature movements (using features at non-rigid locations). These likelihoods were then input to an SVM classifier to recognize the isolated grammatical markers. This yielded an accuracy of 91.76%. We also used our tracker and recognition scheme to recognize the six universal expressions using the CMU databse, and ob-

tained 80.9% accuracy.

While this is a significant milestone in recognizing grammatical markers (or in general recognizing facial expressions in the presence of concurrent head motion), the ultimate goal is to recognize grammatical markers in continuously signed sentences. In the latter problem, simultaneous segmentation and recognition is necessary. The problem is made more difficult due to the presence of coarticulation effects and movement epenthesis (extra movement that is present from the ending location of previous sign to the beginning of next sign). Here, we propose to use the discriminative framework provided by Condition Random Field (CRF) models. Experiments yielded precision and recall rates of 94.19% and 81.36%, respectively. In comparison, the scheme using single-layer CRF model yielded precision and recall rates of 84.39% and 52.33%, and the scheme using layered HMM model yielded precision and recall rates of 32.72% and 84.06% respectively.

In summary, we have advanced the state of the art in facial expression recognition by considering this problem with concurrent head motion. Besides its utility in sign language analysis, the proposed methods will also be useful for recognizing facial expressions in unstructured environments.

List of Tables

List of Figures

List of Abbreviations

AAM Active Appearance Model

AS Assertion

ASL American Sign Language

AU Action Unit

CRF Conditional Random Field

DBN Dynamic Bayesian network

FACS Facial Action Coding System

HMM Hidden Markov Models

KLT Kanade-Lucas-Tomasi

MAT Manually annotated tracks

NEG Negation

NMS Non-manual sign

PCA Principal Component Analysis

PDM Point Distribution Model

PPCA Probabilistic Principal Component Analysis

RH Rhetorical question

SL Sign language

SVM Support Vector Machine

TP Topic

UN Unidentified expressions

WH Wh question

YN Yes/no question

Chapter 1

Introduction

1.1 Sign Language Communication

The deaf communicate through sign language which is a visual-gestural language. Sign languages are used by deaf communities all over the world, with each community usually has its own variation of signing which arises from imitating activities, describing objects, fingerspelling, or making iconic and symbolic gestures. The signs are expressed using hand gestures, facial expressions, head motions and body movements. These visual signals can be cooperatively used at the same time to convey as much information as speech.

When people using different sign languages communicate, the communication is much easier than when people use different spoken languages. However, sign language is not universal, with different countries practising variations of sign language: Chinese, French, British, American, etc. American Sign Language (ASL) is the sign language used in the United States, most of Canada, and also Singapore. ASL is also commonly used

as a standard for evaluating algorithms by sign language recognition researchers.

Many research works show that ASL is not different from spoken languages [1]. The similarities have been found in structures and operations in the signer's brain, in the way the language is acquired, and in the linguistic structure. All languages have two components: symbolic and grammatical components [5]. Symbols represents concepts, and grammatical components provide the way to combine symbols together to encode or decode information. In natural languages, the corresponding analogy is words and grammar; in programming languages, it is keywords and syntax. ASL has both symbolic and grammatical components [5], where, symbols are conveyed by hand gestures (manual channel), and grammatical signals are expressed by facial expressions, and head and body movements (non-manual channel) [5, 1].

For example, consider the sentence

- **English**: Are you hungry?
- **American Sign Language (ASL)**: YOU $[HUNGRY]_{YN}$

In the notation of the above example, *YN* stands for the facial expression of the "yes/no" question; $[HUNGRY]_{YN}$ indicates that the facial expression for the yes/no question occurs simultaneously with the manual sign for *hungry*. This expression is basically formed by thrusting the head forward, widening the eyes, and raising the eyebrows. Without such non-manual signals, the same sequence of hand gestures can be interpreted differently. For example, with the hand signs for [BOOK] and [WHERE], a couple of sentences can be framed as

- $[BOOK]_{TP}$ $[WHERE]_{WH}$ $\rightarrow$ Where is the book?
- $[BOOK]_{TP}$ $[WHERE]_{RH}$ $\rightarrow$ I know where the book is!

The subscripts TP, WH and RH on the words BOOK and WHERE indicate grammatical non-manual signals conveyed by facial feature movements and head motions. The facial gesture for Topic (TP) is used to convey that BOOK is the topic of the sentence. The word WHERE accompanied by a WH facial expression signals a "where?". The hand sign for WHERE made concurrently with the facial gesture for RH indicates the rhetorical nature of the second sentence. When we speak or write, words appear sequentially; i.e., natural languages transfer information linearly. However, our eyes can perceive many visual signals at the same time. Thus the manual and non-manual channels of sign language can be simultaneously used to express ideas.

1.2 Manual Signs

Manual signs or hand gestures, are made from combinations of four basic elements: hand shapes, palm orientations, hand movements, and hand locations. Each of these elements is claimed to have a limited number of categories, for example: 30 hand shapes, 8 palm orientations, 40 movement trajectories, and 20 locations [64].

Signs are created to be visually convenient. During conversation, the Addressee, who is "listening" by watching, looks at the face of the Signer, who is "talking" by signing. Thus, signs are often made in the area around the face so that they are easily seen by the Addressee. From 606 randomly chosen signs, there are 465 signs which are performed near the face area

(head, face, neck locations), and only 141 signs in the area from shoulder to waist [5]. This suggests potential occlusion problems when working with face videos.

Besides, an ASL sentence is also constructed to be suitable for perception by the human visual system. ASL tends to use 3D space as a medium to express the relationship between elements, which can be places, people or things, in a sentence, or even a discourse [1]. At first, the element will be established in space by pointing at some location. This location will later be pointed to when the Signer wants to refer to the corresponding element. Time is also represented spatially in ASL. Space in front of the body represents the future, the right front of the body represents present time, and space at the back represents the past.

The visual characteristic of ASL heavily influences on its grammar. In English, the order of words in a sentence is very important because it decides the grammatical role (subject, object, verb, ...) of symbols, for example:

$$[Peter]_{subject} \text{ likes } [Mary]_{object}$$

However, ASL does not depend on word order to show the relationship among signs. Using 3D space and non-manual signals, ASL can naturally illustrate roles of symbols in a sentence, a paragraph, or a conversation:

Example 1:

$$[P-E-T-E-R-rt]\ \textit{peter}\text{-LIKE-}\textit{lf}\ [M-A-R-Y-lf],$$

Example 2:

$$[M-A-R-Y-lf]^t,\ [P-E-T-E-R-rt]\ \textit{peter}\text{-LIKE-}\textit{mary}$$

In Example 1, the name "Peter" is fingerspelled on the right side. Then, the verb "like" is signed at the middle. After that, the signer points to the left (this sign is denoted by *lf* after the word "LIKE"). Finally, the name "Mary" is fingerspelled on the left side.

In Example 2, the name "Mary" is fingerspelled on the left side together with a topic expression, which is indicated by the small "t" above the name. The comma represents a pause. Following this, the name "Peter" is fingerspelled on the right side, and finally, the verb "like" is signed.

1.3 Non-Manual Signs (NMS)

Linguistic research starting in the 1970's discovered the importance of the non-manual channel in ASL. Researchers have found that non-manual signs not only play the role of modifiers (such as adverbs) but also the role of grammatical markers to differentiate sentence types like questions or negation. Besides, this channel can also be used to show feelings along with signs, as a form of visual intonation analogous to vocal pitch in spoken languages. Non-manual signals arise from face, head and body:

- Facial expressions: eyelids (raise, squint, ...), eyebrows (raise, lower), eye gaze, cheek (puff, suck, ...), lip (pucker, tighten, ...).
- Head motion: turn left, turn right, move up, move down, ...
- Body movements: forward, backward, ...

Bridges and Metzger [15] mentioned six types of non-manual signals used in sign language:

Reflected universal expressions of emotion: the Signer can express one of the universal expressions (angry, disgust, sad, happy, fear, surprise) as his own feeling or somebody else's feeling which he is referring to.

Constructed action: the Signer imitates action and dialog of others from another time or place. For example, when telling a story, the Signer can mimic action in the story.

Conversation regulators: the Signer uses some techniques, usually eye contact or eye gaze, to confirm who he is addressing when there is a group of people.

Grammatical markers: the Signer uses expressions to confirm the type of sentence, or the role of an element.

Modifiers: the Signer uses expressions to add in the quality or quantity to the meaning of a sign.

Lexical mouthing: the signer uses mouth to replace hands for specific signs.

These expressions can be classified into three general types:

- Unstructured expressions: includes reflected expressions and constructed actions. These non-manual signs are used to describe expressions and actions from the past that the signer wants to repeat during a conversation. These expressions do not play a formal linguistic role.

- Lexical expressions: includes lexical mouthing which occurs either with a particular sign, or in place of that sign in a sentence.

- Linguistic expressions: includes conversation regulators, grammatical markers, and modifiers. These non-manual signs provide grammatical and semantic information for the signed sentence.

Since linguistic expressions are non-manual signs that are directly involved in the construction of signed sentences, their recognition is important for computed-based understanding of sign language, and hence they are described in more detail in the following sections.

1.4 Linguistic Expressions in Sign Language

1.4.1 Conversation Regulators

In ASL, specific locations in the signing space (around the signer) called phi-features are used to refer to particular objects or persons during a conversation. While signers use eye contact to refer to people they are talking to, they usually use head tilt and eye gaze to mark object or subject agreements in the signed sentence. This non-manual agreement marking commonly occurs right before the manually signed verb phrase [4].

For example:

Sign: YOU^t *eye gaze to another person* LIKE.

English: He/She likes you.

In the above example, the eye gaze plays the role of she/he in the sentence.

1.4.2 Grammatical Markers

According to [1] and [5], there are eight types of non-manual markers which convey critical syntactic information together with hand signs.

Wh-question: questions that cannot be answered by 'yes' or 'no'; this marker is performed by lowered brows, squinted eyes, tilted or forward head.

Yes/no question: questions that can be answered as 'yes' or 'no'; this marker consists of raised brows, widened eyes, and head thrust forward.

Rhetorical question: questions that need not be answered; marked by raised brows and tilted or turned head.

Topic: topic marker usually appears at the beginning of the signed sentence, or its subordinate clause; consists of raised brows, and single head nod or backward tilt of the head.

Relative clause: Relative clause is used to identify particular things, events or people that the Signer wants to mention. Relative clause marker occurs with all the signs in the relative clause; consists of raised brows, raised cheek and upper lip, and a backward tilt of the head. However, this expression is not common in ASL ([5] page 163).

Negation: negation marker confirms negative sentence; consists of side-to-side head shake and optional lowered brows.

Assertion: assertion marker confirms an affirmative sentence and consists of head nods.

Condition: This type of sentence has two parts: the first part declares the situation, the second part describes the consequence. There are two different markers for the two parts: raised brows and tilted head for

the first part, a pause in the middle, and lowered brows and tilted head in a different direction.

1.4.3 Modifiers

Mouthing is usually used in ASL to modify manual signs. Certain identified mouthings are listed in [15]. Each mouthing type has a certain meaning that is associated with particular manual signs.

For example [15]:

- *Type*: MM.
- *Description*: lips pressed together.
- *Link with*: verbs like DRIVE, LOOK, SHOP, WRITE, and GOING-STEADY.
- *Meaning*: something happening normally or regularly.

1.5 Motivation

Our literature review in Chapter 2 shows that most current works in recognizing facial expressions have focused on recognizing the six universal facial expressions under restrictive assumptions. The common assumptions of these works are isolated expressions, frontal face, and little head motion. These assumptions are inappropriate in the sign language context where the multiple non-manual signs in a signed sentence are usually shown by facial expressions concurrently with head motions. Thus, the recognition of non-manual signs in sign language will extend the current works in facial expression recognition.

Moreover, as extensively reviewed in [85] and Chapter 2, most of the current works on sign language recognition focus on recognizing manual signs while ignoring non-manual signs, with recent exceptions being [108, 79]. Without recognizing non-manual signs, the best system that could perfectly recognize manual signs still would not be able to reconstruct the signed sentence without ambiguity. A system that can recognize NMS will bridge the gap between the current state-of-the-art in manual sign recognition and its practical applications for facilitating communication with the deaf.

In this thesis, we address the challenge of recognizing NMS in sign language and propose schemes for tracking facial features, and recognizing isolated facial expression as well as continuous facial expression. Our focus has been on recognizing six grammatical markers: Assertion, Negation, Rhetorical question, Topic, Wh-question, and Yes/no-question. These grammatical markers have been chosen because they are commonly used to convey the structure of simple signed sentences and deserve to be the next target of sign language recognition after hand sign recognition.

1.5.1 Tracking Facial Feature

Facial expressions in sign language are performed simultaneously with head motions and hand signs. The dynamic head pose and potential occlusions of the face caused by the hand during signing require a robust method for tracking facial information. Based on the analysis in Chapter 2, we propose to track facial features and derive suitable descriptions from them for facial gesture recognition. However, methods like the Kanade-Lucas-Tomasi (KLT) tracker, which are based on intensity matching between consecutive

frames, are vulnerable to fast head motions and temporary occlusions. In Chapter 3, we propose a novel method for robustly tracking facial features using a combination of shape constraints learned by Probabilistic Principal Component Analysis (PPCA) , frame-based matching, and a Bayesian framework. This method has shown robust performance against eye blinks, motion blurs, fast head pose changes, and temporary occlusions.

1.5.2 Recognizing Isolated Grammatical Markers

As described above, grammatical markers are a subset of facial expressions in sign language and consist of facial feature movements and head motions. These two channels have been observed in our data to be uncorrelated and somewhat asynchronous. To address this problem, in Chapter 3, we propose a framework which combines multi-channel Hidden Markov Models (HMM) and a Support Vector Machine (SVM) . This framework analyzes facial feature movements and head motions separately using HMMs and deduces the grammatical marker using an SVM classifier.

1.5.3 Recognizing Continuous Grammatical Markers

Even in a simple signed sentence, multiple grammatical markers appear continuously in sequence. As explained in Chapter 4, beside asynchronization effect between head motions and facial feature movements, continuous grammatical marker recognition also needs to deal with movement epenthesis and co-articulation which affect the appearance of grammatical markers and create unidentified expressions between them. This presents a difficult scenario for generative models such as HMMs. In Chapter 4, we propose a layered Conditional Random Field (CRF) framework which is discrimina-

tive for recognizing continuous grammatical markers. This scheme includes two CRF layers, the first layer to model head motions and the second layer to model grammatical markers. Decomposing the recognition into layers has shown better results than with a single layer.

1.6 Thesis Organization

The rest of the thesis is organized as follow. Chapter 2 provides a literature review of works on facial expression recognition and concludes with a motivation for developing new methods for extracting features and recognizing facial expressions, the essential part of non-manual signs, in sign language. Chapter 3 presents our algorithms for robustly tracking facial features in the presence of head motions and occlusions, and a method for recognizing six common isolated grammatical markers: Assertion, Negation, Rhetorical question, Topic, Wh question, and Yes/No question. This recognition method is also generalized and tested on the six universal facial expressions. Chapter 4 presents our method for recognizing continuously signed grammatical markers (or grammatical marker chains). Chapter 5 completes the thesis with conclusions and possibilities for future works.

Chapter 2

Background

2.1 Facial Expression Analysis

A facial expression is made by movement of facial muscles. Darwin [30] suggested that many facial expressions in humans, and also animals, were universal and had instinctive or inherited relationships with certain states of the mind. Following Darwin's work, Ekman and Friesen [35] found six emotions having universal facial expressions: anger, happiness, surprise, disgust, sadness, and fear. These findings motivated many studies on recognizing facial expressions, especially the six universal emotions, using computer.

Currently, there are many useful applications for facial expression recognition, such as: image understanding, video-indexing, virtual reality, etc. Automatic facial expression analysis methods exploit appearances of human face, using facial textures, and locations, shapes, and movements of facial features to recognize expressions. The relationship between a facial expression and its appearance on a face can be coded by human experts

using some facial coding system like FACS [37] or MPEG4-SNHC [58], or it can be learned by a computer from images.

Ekman and Friesen were interested in the relationship between muscle contractions and facial appearance changes. They proposed the Facial Action Coding System (FACS) [37] for representing and describing facial expressions. FACS includes definitions and methods for detecting and scoring 64 Action Units (AU) which are observable changes in facial textures and head pose. Due to the usefulness of FACS in coding and identifying facial expressions, many efforts are being made to recognize AUs automatically, e.g. [6, 65, 88, 62]. Commonly, a subset of AUs are chosen for recognition. In the training phase, certified FACS experts are required for coding AUs in training images. To overcome differing coding decisions caused by human observations, some agreement among these FACS experts is usually needed. In the testing phase, AUs in each image are recognized, and they are combined to identify the facial expression.

There are many works which analyze facial information. These works can be categorized into: image-based approaches, model-based approaches, and motion-based approaches. Image-based approaches [9, 88] make use of pixel intensities to recognize facial expressions. Tasks in this approach involve facial feature detection, and identifying changes in intensities compared with the neutral expression. The image can be filtered, for example, using Gabor wavelets which have responses similar to cells in the primary visual cortex [42]. Model-based works utilize face models to capture changes on the face. These models are built using the exterior facial structure [3, 23, 17, 44, 39], or internal muscle structure [99]. During an expression, a model-based system tries to deform the model to match with facial

features being observed, possibly using a predefined set of deformations. The matched model is then used to classify the expression. Motion-based facial expression analysis research exploits motion cues to recognize expressions. These motion cues can be obtained by computing dense optical flow or tracking markers on a face in a video sequence [13, 62, 53]. Here, Hidden Markov Models (HMMs) are usually used to recognize facial expressions from motion features.

2.1.1 Image Analysis

Image-based methods utilize appearance information to analyze facial expressions on face images. There are two general approaches: local and holistic. Works following the holistic approach consider face images as a whole. Each n-pixel face image is regarded as a point in n-dimensional space, and face images in training data will form a cluster in high-dimensional space. Statistical methods like Principal Component Analysis (PCA) [27] or Independent Component Analysis (ICA) [6] are commonly chosen to analyze the training data to find subspaces for expressions. A new face image can then be projected into all subspaces, and the nearest subspace can be found to assign the test image to the corresponding expression. A common method used to preprocess face images is to compute the difference image from the peak expressive image and the neutral image of the same person. Another common and effective method is to filter the peak expressive image with Gabor filters which are considered to have similar response properties to cortical cells [42]. Using similar analysis methods as the holistic approach, works using local approach try to apply them on local parts of the face instead of the whole face to avoid sensitivity to identity of person [86].

PCA is used to obtain second-order dependencies among pixels in the image. Applying PCA on a data set of face images gives a set of ghost-like face images called "eigenfaces" [106] or "holons" [27] which are principal components, or axes, of that data set. Any face image can be represented as a linear combination of these principal components. When an image is represented using PCA, it is approximated by projecting to and reconstructing from a space spanned by these axes. After representation by PCA, a face image can be used for person identification or facial expression recognition using recognition methods like nearest neighbors [106], linear discriminant analysis [16], or neural networks [27]. This approach requires high standardization of face images, as any differences in head pose, lighting, or expressive intensity can cause a wrong classification. Calder et al. [16] did a comparison between two approaches for recognizing six universal emotions using two types of preprocessed input data: full-image and shape-free data. Full-image data had been preprocessed so that all face images had the same eye positions and the same distance between eyes. To form shape-free data, input face images were warped to the same average face shape so that facial features were located at standard positions. The approach using full-image data obtained 67% recognition rate while the other achieved 95%. The large difference between these two approaches may come from the higher correspondence among facial features in face images of the shape-free data set.

Bartlett [6, 9] proposed holistically analyzing faces using ICA. Her method aims to separate statistically independent components using information maximization approach. Bartlett stated that ICA can capture the high-order statistical relationship among pixels, while PCA can only cap-

ture the second-order relationship. Moreover, she also mentioned that high-order statistics captured the phase spectrum of the image which was more informative than amplitude spectrum captured by second-order statistics. Data used in Bartlett's work was frontal face images which were cropped, centered, and normalized. Locations of eyes and mouth were used as references for centering and cropping. Neural networks were used for unsupervised learning of ICA parameters. Bartlett reported that her system was able to recognize 12 Action Units with 95% accuracy which was claimed to be better than recognition rates of both naive and expert humans.

Further, Barlett et al. [66] presented detailed comparative results for recognizing the six universal expressions with various types and combinations of classifiers. Though the database consisted of frontal face videos, the experiments were performed on the peak expressive frames. The best recognition accuracy of 93.8% was obtained with an RBF kernel SVM, with optimal Gabor features selected by Adaboost. The classifier was applied on video sequences for classifying each frame. The 7-way classifier outputs (including the neutral expression) plotted as a function of time were found to closely match the expression that appeared in the video. Generalization to an unseen dataset lowered the accuracy to 60%, suggesting that a large training corpus may be needed to generalize across different environments. Moreover, pose variations were not considered.

Padgett and Cottrell [86] compared different feature representations: whole face image, local patches at main facial features (mouth and eyes), and local patches at random locations on the face. As with Cottrell's previous work [27], they used PCA on these features and performed classification using neural networks. They found that the representation using local ran-

dom patches obtained 86% recognition rate which was better than local patches (80%) and whole face image (72%). However, their experiment was based on manually locating facial features on the face. When facial features were manually located approriately, the feature representation became almost noise-free which might be the reason for the good classification result of local patch-based representations. Donato et al. [33] also reported that there was hardly any difference in recognition result between holistic and local features.

Gabor wavelet filters [31] can extract specific spatial frequency and orientation by using a Gaussian function modulated by a sinusoid. Gabor filters can be used to preprocess face images to remove most of the variabilities due to lighting changes and reveal local spatial characteristics of facial features. Bartlett [6] claimed that face images filtered using Gabor wavelets gave outputs similar to ICA, and both representations led to high facial expression recognition rate, of more than 90% [9, 70].

Pantic [88, 87] followed the local approach, though feature representation in these works was based on geometrical characteristics of facial features instead of pixel-based statistics or Gabor wavelet responses. Her work aimed to recognize all 44 Action Units using frontal and profile images. Pantic heavily relied on facial feature detectors to locate facial features on neutral and expressive face images. Geometrical measurements were performed on facial features and a rule-based classifier was used to identify Action Units. Then another rule-based classifier was used to recognize the six universal emotions using the recognized Action Units. This method may not be able to deal with natural head motions because it will be difficult to correctly locate facial features.

Image-based facial expression analysis works usually use static and standardized face images. Extracting features is not a big challenge with this approach. However, image-based methods are highly sensitive to head pose and do not consider temporal characteristics of facial expressions for recognition.

2.1.2 Model-based Analysis

Along with using pixel intensities of face images, model-based facial expression analysis works also exploit face shape and structural constraints. In this approach, the first task is to build the face model. Face models can be 3D meshes [99, 39, 32, 21], 2D meshes [23, 59], 2D point distribution models [51], etc. These models can be deformed using physical parameters [32, 59, 21], anatomical structures [99, 39], principal shape and texture components [23]. Face models are usually used to track a face in a video sequence and capture its expressions, so initializing a model on a face image becomes the next significant task. Many works currently rely on manual initialization to initially align the model, even though there are many methods to automatically detect the face [93, 114, 107] and locate facial features [29, 74, 43]. Faces and facial features are tracked using active contours [99], image templates [21], optical flow [39, 32], or linear regression computations on matching errors between the model and the face image [23]. Tracking results are then utilized to create parameters for deforming the model. Deformations of face models are later employed to analyze or synthesize facial expressions.

Terzopoulos and Waters [99] combined physically-based 3D mesh with anatomically-based facial control process to form a realistic 3D dynamic

model of the face, which had three layers to simulate muscle, dermis and skin tissue layers. The final model had 6 representation levels: images, geometry, physics, muscles, control and expression. To express an emotion (expression level), corresponding muscles (muscle level) were stimulated by an activating mechanism (control level) using predefined knowledge, through a simplified form of FACS; contractions of simulated muscles deformed the simulated dermis layer physically (physics level); deformations at dermis layer caused distortions on the geometrical mesh simulating skin tissue (geometry level); the model's surface was rendered from these distortions to form the output appearance (image level). To learn control parameters for the model, facial expressions were analyzed using active contours. Human subjects were heavily made up to intensify nine high gradient facial contours including hairline, eyebrows, nasolabial furrows, tip of the nose, upper and lower lips, and chin. Active contours, or snakes [57], were manually initialized and used to track these intensified facial features over a video sequence of the subject's performance of a required expression. Non-rigrid shapes and motions of contours provided quantitative information to compute parameters used to rescale the model and rebuild the expression. The authors claimed that the analyze-and-synthesize process could be done in real-time. There are also some drawbacks to this work. Firstly, heavy make up and manual initialization are required to help snakes track better. Secondly, the system works with frontal face and static head only, and there is no guarantee that snakes will appropriately work with natural head motions which cause 3D movements of facial features. Besides, a lot of work is required to fully construct muscles on the model.

Essa et al.[38, 39] also used a geometrical, physical, anatomical, and

control-based dynamic model to synthesize and analyze the six universal expressions. The model, which had only one layer, was built using finite elements and could simulate not only the stiffness and the damping but also the inertia which was missing from Terzopoulous's model. Simoncelli's optical flow estimation method [95] was used to analyze facial expressions. In each frame of a video sequence containing a facial expression, dense optical flows were computed at every pixel. The face image in each frame was divided into 80 regions, and the flow in each region was averaged and located at its centroid. The synthesis process accepted this optical flow as input, and a feedback loop employing Kalman filter was used to obtain parameters, considered as muscle actuations, to optimally deform the model. The movement of chosen shape control points on the model was called FACS+, i.e. FACS with temporal information. This work also required frontal view of the face and static head to correctly compute dense optical flows, and required heavy computations. In an effort to make the system work in real time, the author used image matching instead of optical flow to compute deformation parameters. At first, normalized peak expressive images for expressions and corresponding deformation parameters are stored. With each frame, the smallest difference value between the stored expressive images and the current frame was obtained. This difference value was fed into a RBF network to find the corresponding deformation parameters. These parameters were optimized using a framework based on Kalman filter. Similar to works using the image-based approach, this modification relied on particular face pose, was person dependent, and assumed static head.

Cohen et al. [21] used the Piecewise Bezier Volume Deformation (PBVD) tracker developed by Tao and Huang [98] for face tracking and feature ex-

traction. A 3D model used by the PBVD tracker was built using the finite element method and owned physical (but not anatomical) characteristics like Essa's. The model was composed of 16 planar patches connected by hinges, and each patch was modeled as a polygonal mesh resembling an elastic membrane. The deformation of each patch could be done by a linear combination of vibration modes defined to maintain the smoothness of patches and low computational cost. In the tracking stage, salient facial feature points were manually chosen in the first frame of a video sequence to initialize the model. Nodes of each mesh were tracked using an image matching method. After that, weighted parameters for vibration modes were estimated using least squares method to minimize the difference between the deformation of the patch and nodal displacements. Recovered motions were used to form Motion Units which were motion vectors containing numeric magnitudes of predefined motions of facial features. Motion Units were claimed to represent not only motions of facial features but also the intensity and the direction of the motion. Motion Units were used both to recognize the six emotional universal expression and to segment these expressions which are continuously recorded in a video sequence [21]. The PBVD tracker worked well with in-plane but not with out-of-plane movements [98].

A 2D elastic mesh called Potential Net was used by Kimura [59] to recognize three expressions: happy, anger, and surprise. The mesh was a rectangular grid, where each node was connected to four other nodes by simulated springs. Nodes on the boundary were fixed, while interior nodes could be moved by combined forces from elastic springs and gradients of the image. In each frame of a video sequence, the face and facial features were

manually detected, the face area was then extracted and normalized; there is also an effort to automatically detect the face area using the Potential Net itself [11]. Differential filter and Gaussian filter were sequentially applied on the face area. After alignment on the face area, the Potential Net will be deformed by the force computed from the image gradient and the internal elastic force. Motion vectors formed from displacements of nodes are used for later classification. However, the author just reported a simple investigation of feature vectors. It aappears difficult to extend this kind of model to cope with head motions because it relies on frontal view and 2D mesh.

Instead of using elastic models, Cootes [24] proposed the Point Distribution Model (PDM) which can both represent typical shape of an object and permit variability. The model was built from a training image data set which represented varying shapes of an object. At first, in each image, a set of labeled points was marked along edges best representing the object. The mean shape of the object and its deviations were then computed from these training sets to form training shapes. Principal component analysis was applied on these training shapes to find main modes of shape variations. Deformations of the model were later done by adding a linear combination of main modes to the mean shape. Parameters associated with main modes were also interpreted as shape control parameters. During tracking of the object in a video sequence, shape control parameters can be iteratively adjusted to minimize the error computed by some matching function. PDM can be used to track face and facial features, and parameters found in tracking can be used to classify facial expressions such as the six universal emotions [51]. Head motions were required to be minor to avoid 3D

distortions of facial features.

The Active Appearance Model (AAM) suggested by Cootes [23] was a more extensive version of PDM which combined both shape model and texture model. Like building a shape model, a texture model was also built from training image data. Mean gray-level texture was obtained, and main modes of gray-level texture were learned. New texture was then synthesized by adding a linear combination of main texture modes to the mean texture. The search process with AAM aims to reduce error between synthesized 2D face image and the input image. Much effort is being made to overcome drawbacks of AAM like limited head motions [34, 110], occlusions [46], person dependence [45], etc. Cristinacce and Cootes [28] propose an automatic template selection method for facial feature detection and tracking. This uses a PCA-based shape model and a set of feature templates learned from training face images. During tracking, the method iteratively selects a set of local feature templates to fit an image, while constraining the search by the global shape model.

In general, model-based works follow an analysis-by-synthesis scheme. The learned models have constrained variances which helps the classification of certain expressions with less ambiguity. However, most of the works focus on recognizing six universal expressions with frontal view, and static head or with minor head motions. None of them makes an effort to identify facial expressions occurring with natural head motions.

2.1.3 Motion Analysis

Motion-based works try to detect and analyze facial expressions based on analyzing movements of face pixels in consecutive frames of a video se-

quence. An essential motivation for this approach is based on the work done by Bassili [10] who showed that moving dots on a face provided significant information for emotion recognition. Two common methods in the literature are used to capture motion cues on the face: optical flow [71, 13, 112, 113, 63, 3] or tracking facial features [65, 62, 53, 116, 47].

Mase [71] inspired other researchers by using optical flow to analyze facial expressions on frontal face. He computed dense optical flow on video frames to recognize facial muscle actions and recognized four emotions: happiness, anger, disgust, and surprise. At first, a dense optical flow was computed using Horn and Schunck's gradient based algorithm. The author used two recognition approaches based on optical flow. In his top-down approach, a set of windows corresponding to underlying facial muscle structure was then placed on the face, and optical flow field inside each window was averaged and assigned at its center. These averaged optical flow vectors were considered as signatures of muscle movements and were claimed to be related to Action Units. Emotional expressions were identified based on these muscle movements using FACS-based descriptions. In his bottom-up approach, the original dense optical flow was divided into rectangular regions. After that, feature vectors were formed using averaged PCA on the first and second moments of the optical flow fields in each region. K-nearest-neighbor was then used to recognize four emotional expressions. His work did not address problems like head motion and consecutive expressions.

Yacoob and Davis [112, 113] worked toward computing optical flow to analyze feature movements to recognize six universal emotions. The authors aimed to describe basic motions of regions corresponding to facial fea-

tures. At first, facial features (mouth, nose, eyes, eyebrows) were detected and rectangles around these features were located. Next, ways or directions these rectangles deformed during a facial expression were identified by computing optical flow using Abdel-Mottaleb's method [2]. A dictionary for facial dynamics was also developed. Each entry of the dictionary involved three parts: facial component, basic action of that component, and motion cue. A motion cue was identified from the optical flow. Every facial expression was considered to involve three temporal periods: the beginning, peak and ending. A facial expression was recognized from the basic actions of facial components in corresponding temporal parts. The work was done on video sequences of frontal faces and static heads.

Black and Yacoob's work [13] was an improvement of Yacoob and Davis' work above. Black also identified temporal moments of facial expressions by optical flow computation. Emotional expressions were recognized based on a facial motion dictionary. There were two developments in Black's work. First, optical flows of non-rigid facial features, mouth and eyes, were computed separately from the rest of the face. This separation provided a way to differentiate between non-rigid and rigid motions on face. Second, optical flows were characterized by affine parameters. It helped to capture better the facial motions caused by non-rigid movements and 3D head motions. Regions of non-rigid facial features for computing optical flow was also deformed based on computed optical flow's affine parameters. The final system could recognize local facial feature movements, six emotions, and 14 head motions (rightward, leftward, upward, downward, expansion, contraction, horizontal deformation, vertical deformation, clockwise rotation, counter clockwise rotation, rotate right about neck, rotate left about

neck, rotate forward, rotate backward). Planar assumption of the face to use affine parameters, heavy computation for dense optical flow, and the sensitiveness of optical flow computation with lighting changes and occlusions are drawbacks of this work.

Anderson [3] and Liao [63] are recent works which use optical flow for recognizing facial expressions. They both analyzed optical flow on local facial regions to recognize six emotions. While Anderson focused on planar analysis, Liao approximated the head as a 3D cylinder to estimate 3D motions. Anderson's was optimized on the image size and the number of frames precessed per second to work in real time while natural head motions was not considered. In Liao's work, 3D head pose was recognized using Xiao's approach [111] to remove the effect of head movement in optical flow computation. Optical flow was described by affine parameters. Local regions were defined on the surface of the 3D cylinder, and a predefined interdependence among regions was used in the classification phase. Even though his system could classify facial expressions with head motions, head motions themselves were not well addressed, nor was occlusion.

Optical flow analysis can capture subtle movements on the whole face, and estimate head motions. However, heavy computation is always a problem with applications using optical flow.

Lien [65] tried and compared three methods for facial expression analysis: image analysis, optical flow analysis, and facial feature points tracking. He analyzed images to capture wrinkles appearing during an expression, computed the optical flow to estimate movements of both smooth or textured regions on face, and tracked facial features to identify facial actions on high texture regions: brows, eyes, nose and mouth, which are also highly

related to muscle activations. His work aimed to recognize three upper and six lower facial expressions described using Action Units. The averaged recognition rates of all three methods were quite high: 92%, 86%, and 83% for dense optical flow analysis, feature points tracking and image analysis, respectively. He found that feature representation from facial feature points tracking was fast, accurate, and could cope with large head motions. Cohn [22] also states that feature point movements are good enough to analyze facial expressions.

Tian [62] used Multi-state Component Models for tracking facial features to recognize Action Units. These models exploited geometrical characteristics of facial features at different states for tracking in cases where one or more feature points were missing due to facial actions like eye blinking or lip sucking. Feature points were manually marked on the first frame around brows, eyes, and mouth. They were automatically tracked over subsequent frames using the KLT algorithm. In each frame, relative positions of feature points were used to estimate current states of facial features. From the estimated state and positions of feature points, appropriate prebuilt 2D models were chosen corresponding to current states of tracked facial features. The input video sequence stopped at the peak of the expression. Parameters measured on facial features' shapes at the first and final frame were fed into feed forward neural networks to identify presented Action Units. Due to the advantage of feature points tracking method, her system can cope with head motions as long as the face is still frontal. However, head motion recognition was not considered.

Kaliouby [53] used a similar method to Tian's to recognize six cognitive mental states (agreement, concentrating, disagreement, interested, think-

ing, unsure). Twenty four facial landmarks were automatically located in the first frame and tracked across the video sequence. Displacements of these landmarks were used with left-to-right HMM to recognize head motions and facial feature movements. His system could recognize four head motions (head nod, head shake, tilt display, turn display) and two facial displays (lip pull, lip pucker) at above 95% recognition rate. Outputs of HMMs were fed into Dynamic Bayesian Networks (DBN) to identify mental states. The approach of this work is quite similar to ours, except that it aims to apply it as an "emotional hearing aid". Tracking techniques are not developed to cope with occlusions or lost facial features. Besides, facial expression segmentation is also not considered.

Ji et al. consider facial feature tracking and expression recognition in [117, 104, 103, 116]. In [117], a set of 28 facial features are automatically detected with Gabor filters and tracked under varying pose and facial expressions using Kalman fillters at the 28 locations. Pose is estimated using feature points at rigid facial locations, the weak perspective model, and a PCA-based shape constraint is used. In [104], a multi-state hierarchical facial feature model is used to handle facial expression changes, with tracking implemented by a Switching Hypothesized Measurements (SHM) filter. 3D pose is estimated from tracked feature points to constrain the feature search. In [103], a mixture PPCA model is proposed to model shape variations due to pose, and constrain the (Gabor) feature matching process during tracking. Here, no dynamics of the transitions between the mixture components is used. In [116], Dynamic Bayesian Networks (DBN) are used for modeling facial expressions from video. IR illumination is used to reliably locate the pupils, after which several feature points on the face

are detected and tracked by Kalman filtering. The motion of the feature points are manually associated with FACS action units and represented by a DBN. Plots of the probabilities of the six universal expressions vs. time indicated good agreement with the facial expressions, even with changing head poses. A useful feature of their method is the integration of temporal information to induce robustness with respect to occlusions.

Feature tracking is a promising approach for analyzing facial expressions occurring with head motions. More work needs to be done to develop a reliable tracking method when there are head motion and occlusions.

2.2 Recognizing Continuous Facial Expressions

In sign language sentences, more than one facial expression may be used together with hand signs. As mentioned in Chapter 1, facial expressions and head motions, which form the non-manual channel, provide linguistic information to the hand sign channel. Segmenting facial expressions captured in a video sequence is necessary for fusing this with information from the manual channel to achieve complete recognition of signed sentences.

Black and Yacoob's work [13] is a pioneering work in recognizing continuous facial expressions with head motion. Affine-like parameters of facial feature movements and head motions were extracted from dense optical flow. Rule-based discriminative models classified facial feature movements and head motions separately. They obtained an average recognition rate of 88% and 73% on laboratory data, and real life data (from television programs), respectively. Their method required a short neutral expression between different facial expressions.

Chang [20] used a low-dimensional manifold for modeling and tracking a face, segmenting and recognizing six facial expressions in video sequences. In the training stage, Active Shape Models were used for detecting and tracking 2D facial landmarks in video sequences. One ASM was built for each type of expression. Shapes of 2D facial features tracked by ASM in each frame were normalized and projected into a low-dimensional manifold using a nonlinear dimensionality reduction method [52] which can maintain the main geometrical structure of the data. In the low-dimensional manifold, projected face shape changes during a facial expression formed a path starting from the center corresponding to the neutral face. In the testing stage, ICondensation method was used to control the tracking process by predicting the deformation of an ASM or choosing another one which better matched with the current facial expression. At every frame, the ASM of the expression being used was considered as the recognized facial expression; the change to another ASM would mark the end of the previous expression. Their system performed well in tracking faces with different expressions in long video sequences. However, head motions were not addressed in this work.

De la Torre et al. [60] proposed a framework for detecting rare facial gestures. Personalized AAM [72] was used for tracking subjects' faces during an interview. The neutral facial gesture was automatically detected by applying spectral clustering on dynamic feature vectors combining shape and appearance. A greedy approach was used for detecting segments of facial behaviors by hierarchically matching with predefined patterns. Quantitative assessment of the detection was not reported.

Hoey [50] considered the problem of unsupervised classification and seg-

mentation of facial expressions in video sequences. A multilevel dynamic Bayesian network (DBN) was used to learn models characterizing facial expressions and their high-level syntactic relationships simultaneously. They worked with video sequences of five emotional expressions appearing in a predefined order: disgust, fear, happy, sad, and surprise. Training and testing data were generated using a simulation model which combined isolated facial expressions in the predefined order, which may be not appropriate in realistic cases. The accuracy was more than 88%, however they were evaluated in a very constrained manner where facial expressions in video sequences were shown in the same order and separated by a short pause.

Cohen et al. [21] used a piecewise 3D wire frame model-based approach developed in [98] for tracking 16 facial features (selected manually in the first video frame), and estimated their 3D motions. These were mapped to "motion units" and used as the basic features in a multi-level HMM scheme for classifying the six universal expressions and the neutral expression. The classifier provides implicit segmentation and recognition of video sequences containing multiple expressions. They reported 82.46% and 58.63% accuracy for person dependent and person independent tests, respectively, on their database of 5 persons. The experimental results were reported on sequences where expressions transited through the neutral expression. The training and testing data were constructed to conform to this constraint.

As generative models, HMMs suffer from two weaknesses: the statistical independence assumption of observations and the difficulty in modeling their complicated underlying distributions. On the other hand, the Conditional Random Field (CRF) proposed by Lafferty et al. [61] is a discriminative model which avoids these weaknesses. Kanaujia and Metaxas [56] used

the CRF to recognize the six universal expressions and obtained promising results. Quattoni et al.[90] proposed Hidden-state CRF (HCRF) models and obtained an accuracy of 85.25% for recognizing head shakes and head nods.

Chang et al. [19] proposed a modified HCRF called Partially-Observed HCRF (PO-HCRF) which allowed observations of hidden states to be assigned to selected frames. It was demonstrated that PO-HCRF performed better than HCRF on recognizing the six universal facial expressions and an SVM-AdaBoost scheme [8] on recognizing 15 Action Units. The PO-HCRF achieved an accuracy of 80.1% with 9.18% false alarm rate for recognizing "continuous" facial expressions in simulated sequences created by concatenating sequences of isolated expressions.

2.3 Recognizing Facial Gestures in Sign Language

As extensively reviewed in [67, 85], most of the current works on recognizing sign language still focus on recognizing manual signs while non-manual sign recognition has by and large neglected. In recent works, Von Agris et al. [109] propose a user adapted AAM model to identify areas of interest such as the eyes and mouth region, and suggest processing steps to estimate head pose, gaze direction and lip outline along with other distances between facial features. A simple scheme is proposed to detect facial occlusions by hand. However, no tracking or detailed non-manual classification results were reported.

Vogler and Goldenstein [108] proposed a tracker based on a 3D de-

formable model. Tracking using these models is sensitive to facial occlusions by the hand during signing, and hence, an outlier rejection mechanism is proposed to deal with the occlusions. Good tracking results during occlusions have been shown. A qualitative comparison of head pose angles extracted from the tracker, with discrete ground truth labels showed good agreement. However, the 3D face model needs to be fitted to each subject, which can be a laborious process.

Recently, Neidle et al. [79] considered recognition of Wh question (*WH*) and negation (*NEG*) facial expressions in ASL signed sentences. The ASM-based tracking scheme proposed in [55] is used to track face and facial feature movements, and to estimate head pose (pitch, yaw, and tilt) in each frame. A video sequence is labeled as either *WH* or not, by classifying each frame and using majority voting. A stacked SVM formed by three SVMs is used to classify each frame. The presence of *WH* expression is evaluated separately by two SVMs based on the appearance of the eye and eyebrow region and the pitch angle of the head. The third SVM is used to confirm the presence of *WH* using the scores output by the other two SVMs. A similar approach is used for the *NEG* expression. They reported recognition accuracies of 100% and 95% for *WH* and *NEG*, respectively.

2.4 Remarks

Most works in the facial expression recognition literature focus on recognizing six universal emotions: happy, anger, sad, disgust, surprise, and fear. Most considered video sequences where the heads are relatively stationary or only static images. Works using model-based approaches may not cope

well with subtle facial feature movements and head motions. Motion-based approaches using frame-based tracking algorithms like KLT are vulnerable to fast eye blinks, head motions and temporary occlusions as shown in Chapter 3. Besides, attempts for recognizing continuous facial expressions usually assume little head motion while head motions are not only required but also an important cue for recognizing facial expressions in sign language. The assumption of a neutral state between facial expressions is also not applicable in sign language where there are natural transitions between facial gestures. Finally, works in recognizing non-manual signs in sign language, are still at an early stage, and we hope that our work will be a useful contribution in this direction.

Chapter 3

Robustly Tracking Facial Features and Recognizing Isolated Grammatical Markers

3.1 Introduction

Many works on facial expression recognition in the literature [88, 41] are not suitable for direct application to sign language, as they commonly assume frontal face, stationary head, and no occlusions, e.g. [27, 7, 88, 16]. Dense optical flow analysis was used for identifying facial expressions and head motions [13] but this approach is computationally heavy and is sensitive to fast head motions and occlusions. As the movements of facial features are regulated by facial muscles and communicative customs [36], the face shape and its deformations during an expression can be modeled. Tracking facial features using face models provides the flexibility of representing facial feature movements with head motions and robustness to noise. 3D

models [100, 39, 32, 21] provide a mechanism for estimating head pose. However, they are computationally intensive for tracking and adaptation. 2D models [115, 25, 62] are simpler in this respect and many works have been presented to cope with different head poses by using multiple linear models modeled by Principal Component Analysis (PCA) [92], using non-linear 2D models [20], or by combining with a 3D model [105].

In this chapter, we address the problem of tracking facial features robustly and recognizing isolated facial expressions in ASL. We propose and investigate the performance of two algorithms for tracking facial features exhibiting facial expressions, possibly with concurrent head notion, and occlusion, using spatio-temporal shape constraints. These constraints are provided by a learned mixture of Probabilistic Principal Component Analysis (PPCA) [102] model and integrated with recursive tracking schemes. In one scheme, a textural match measure is optimized in every frame with a penalty term for face shape deviation from a recursively predicted PPCA subspace. In the other scheme, observations obtained from a Kanade-Lucas-Tomasi (KLT) tracker [69] are refined by projection and reconstruction from a recursively predicted PPCA subspace. An update scheme called Incremental PPCA suggested in [81] is adopted to improve the robustness of tracking to face shapes of different people.

For recognizing facial expressions in ASL, appropriate distance measures are derived from tracked facial features to minimize the effects of head motion, and are input to a set of Hidden Markov Models (HMM) to evaluate the likelihoods of characteristic facial feature motions. The likelihoods of head motions are evaluated by another set of HMMs using motion vectors of facial features at non-deformable facial feature locations. These

likelihoods are all input to a Support Vector Machine (SVM) to identify six common grammatical expressions: *Yes/no question* (*YN*), *Wh question* (*WH*), *Topic* (*TP*), *Negation* (*NEG*), *Assertion* (*AS*), and *Rhetorical* (*RH*). Since the *conditional clause* marker has complex structure and the *relative clause* marker is uncommon in ASL, we will consider these markers in our future works.

In the following, in Section 3.2, we briefly describe PPCA and develop two algorithms for robust tracking. In Section 3.3, we consider expression recognition using the tracked features and describe the features derived for input to a set of multichannel HMMs, whose output likelihoods are then used in an SVM for recognition. Section 3.4 gives extensive experimental results, comparisons and discussion, and Section 3.5 concludes this chapter.

3.2 Robust Facial Feature Tracking

Facial expressions in sign language occur concurrently with head pose changes and hand signs. Head motion can be quite fast resulting in motion blur in the acquired video, and the face can also be occluded by the hands during signing. This is a challenging situation for tracking facial features; a simple tracker based only on differences between adjacent frames can easily drift away from the facial features. A robust tracker for this scenario needs to be constrained appropriately; a natural constraint is to require that the tracked feature points conform to a model of face shape. We propose to model face shape by learning a mixture PPCA model from training video. The mixture components or subspaces represent homogeneous clusters of head pose and facial expressions. The advantage of PPCA is the probabilis-

tic interpretation that it allows for the PCA subspaces, and for evaluating the likelihood of face shapes. During training, the dynamics of the face shape transitions between subspaces are also learned.

Based on the learned face shape and transition models, we propose and experiment with two facial feature tracking algorithms. In Algorithm 1, we use the learned face shape transition dynamics to predict the face shape subspace at the next time instant. An iterative optimization scheme is then used to minimize an objective function which consists of a match measure for the feature points and a penalty term for face shape deviation from the predicted face subspace. In Algorithm 2, we incorporate the KLT tracker [69] into a recursive Bayesian scheme, which also uses the learned face shape model and transition dynamics. Though the KLT algorithm works well in simple situations, natural head motions and temporary facial occlusions are inevitable in sign language communication. In such situations, gradients in the vicinity of tracked feature points can change abruptly, causing the KLT algorithm to track incorrectly. Thus we use the KLT tracker to provide raw observations for track propagation and smooth it in the Bayesian scheme to be consistent with face shape.

3.2.1 Construction of Face Shape Subspaces

We chose the $N = 21$ facial feature points on the eyebrows, eyes, nose, and mouth as shown in Fig. 3.1 to represent face shapes and classify facial expressions. Minor variations of these points have been used in the literature, but we have found this set of feature points to be useful to discriminate among the sign language (SL) expressions of interest as well as the six universal expressions. The eye corners and the points around the nose are good

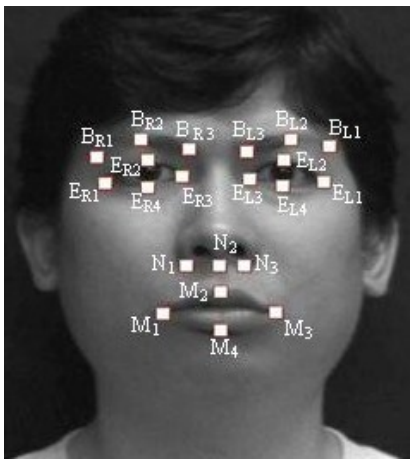


Figure 3.1: Feature points of interest.

indicators of rigid head motions, while the others are indicators of facial feature deformation. A small $w \times w$ window of pixels centered on each of these feature points is used for intensity matching. Let face shape in a frame be represented by a vector of the N feature points, $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1^T \quad \tilde{\mathbf{z}}_2^T \quad \ldots \quad \tilde{\mathbf{z}}_N^T]^T$, where $\tilde{\mathbf{z}}_k = [\tilde{x}_k \quad \tilde{y}_k]^T$ represents the coordinates of the kth feature point. Correspondingly, $I(\tilde{\mathbf{z}}_k)$ denotes the vector of concatenated intensity values from the $w \times w$ window centered on $\tilde{\mathbf{z}}_k$ in the image, and $I(\tilde{\mathbf{Z}})$ is a vector formed by stacking $I(\tilde{\mathbf{z}}_k)$ vectors, $I(\tilde{\mathbf{Z}}) = [I(\tilde{\mathbf{z}}_1)^T \quad I(\tilde{\mathbf{z}}_2)^T \quad \ldots \quad I(\tilde{\mathbf{z}}_N)^T]^T$. We manually mark these facial feature points on training video frames to obtain a set of face shapes $\{\tilde{\mathbf{Z}}\}$ and the corresponding intensity vectors $\{I(\tilde{\mathbf{Z}})\}$.

For each face shape $\tilde{\mathbf{Z}}_t$ marked in image space, a normalized face shape $\mathbf{Z}_t$ is obtained by using a similarity transformation,

$$\mathbf{Z}_t = \mathbf{A}_t \tilde{\mathbf{Z}}_t + \mathbf{b}_t \tag{3.1}$$

These normalized training face shapes are grouped into subspaces using

a mixture PPCA model. Our motivation for using PPCA to represent and partition the face shapes is the associated probability density which is lacking in PCA. This provides the likelihood for face shape which we use in our tracking schemes.

The PPCA Model [101]

A Gaussian latent variable model for Z can be written as

$$\mathbf{Z} = \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is the noise model and $\boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{I})$ is a q-dimensional vector of latent variables; $\boldsymbol{\mu}$ is the mean and $\mathbf{W}$ is the $d \times q$ loading matrix relating the d-dimensional observation $\mathbf{Z}$ ($d = 2N$) to the latent variables. This induces a Gaussian density for $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ where the model covariance is given by

$$\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T \tag{3.3}$$

and a Gaussian posterior distribution for the latent variables,

$$p(\boldsymbol{\alpha}|\mathbf{Z}) \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{Z} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M}) \tag{3.4}$$

where $\mathbf{M} = \sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W}$, and $\mathbf{I}$ is the identity matrix.

It is shown in [101] that by maximizing the log-likelihood of the L observations with respect to the model parameters, we obtain:

$$\boldsymbol{\mu} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{Z}_i \tag{3.5}$$

$$\mathbf{W} = \mathbf{U}(\boldsymbol{\Lambda} - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{3.6}$$

where $\mathbf{U}$ is a $d \times q$ matrix whose columns are the q principal eigenvectors of the sample covariance matrix of $\mathbf{Z}$ and $\mathbf{\Lambda}$ is a $q \times q$ diagonal matrix of corresponding eigenvalues; $\mathbf{R}$ is an arbitrary rotation matrix, which we simply choose to be the identity matrix. The noise variance is given by

$$\sigma^2 = \frac{1}{d-q} \sum_{i=q+1}^{d} \lambda_i \tag{3.7}$$

which is the average variance of the discarded dimensions.

Since $\boldsymbol{\alpha}$ is specified by a posterior distribution, an observation $\mathbf{Z}$ can be represented in the latent space by the posterior mean,

$$\bar{\boldsymbol{\alpha}} = \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{Z} - \boldsymbol{\mu}) \tag{3.8}$$

and an optimal reconstruction in normalized face space can be obtained as

$$\hat{\mathbf{Z}} = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{M}\bar{\boldsymbol{\alpha}} + \boldsymbol{\mu} \tag{3.9}$$

with the same reconstruction error as PCA. The optimal reconstruction $\hat{\hat{\mathbf{Z}}}$ in the image frame, or image space, can be obtained from $\hat{\mathbf{Z}}$ by inverting the similarity transformation in Eq. 3.1.

The model can be generalized to a mixture of PPCA as

$$p(\mathbf{Z}) = \sum_{i=1}^{K} \beta_i \rho(\mathbf{Z}|S^i) \tag{3.10}$$

where β_i are the mixing weights, K is the number of mixture components, and $\rho(\mathbf{Z}|S^i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{C}_i)$ is the PPCA model for the ith subspace S^i with

covariance matrix given by

$$\mathbf{C}_i = \sigma_i^2 \mathbf{I} + \mathbf{W}_i \mathbf{W}_i^T \tag{3.11}$$

and where σ_i^2 and $\mathbf{W}_i$ are analogous to σ^2 and $\mathbf{W}$ for the single component case. These parameters and β_i can be estimated by maximizing the log-likelihood using the EM algorithm.

The EM algorithm requires the number of mixture components, K, and the initial conditions for the iterations to be specified. We obtain these by using the G-means algorithm of [49]. Here, starting with an initial number of cluster centers (e.g. one) in the k-means algorithm, the Anderson-Darling statistic [97] is used iteratively to increase the number of clusters until each cluster can be represented by a unimodal Gaussian distribution. In each iteration of the algorithm, hypothesis testing based on the Anderson-Darling statistic is used to verify whether the data assigned to a cluster are samples from a Gaussian distribution; if not, the cluster is split into two sub clusters. The only parameter to be set is the significance level γ for the test, whose choice controls the number of clusters obtained. The clusters thus obtained are used to initialize the EM algorithm to estimate the mixture of PPCA model for the face shapes.

Once the model is learned, the training face shape vectors are hard assigned to a PPCA subspace by the maximum probability rule,

$$\mathbf{Z}_i \in S^k \text{ where } k = \underset{l}{\operatorname{argmax}}\, \rho(\mathbf{Z}^i | S^l) \tag{3.12}$$

This partition of the training shape vectors is useful to learn the subspace transition probabilities for the tracking algorithm.

Characterizing Face Shape Transitions

As the facial expression evolves, the face shape will make transitions between the learned PPCA subspaces. The training data can be used to learn the probabilities of transitions between subspaces for use in the tracking algorithm. For this, the training face shapes are indexed by the subspace k they belong to as in Eq. 3.12, and also according to their time index t in a given video sequence. The transition probability from S^i to S^j is computed as the ratio of the number of $i \rightarrow j$ transitions in consecutive frames over all sequences, to the total number of samples in S^i

$$P(S^j|S^i) = \frac{Count\left(\{\mathbf{Z}_{i,t}, \mathbf{Z}_{j,t+1}\}\right)}{Count\left(\{\mathbf{Z}_{i,t}\}\right)} \tag{3.13}$$

3.2.2 Track Propagation

To estimate the face shape $\tilde{\mathbf{Z}}_t$ in the current frame, the subspace for the current frame is first predicted and the optimal face shape estimate $\tilde{\mathbf{Z}}_{t-1}$ in the previous frame is used as the initial condition for iterative optimization in the predicted subspace as described below. Here, $\tilde{\mathbf{Z}}_t$ is found as an acceptable compromise between the matching of the intensities $I_t(\tilde{\mathbf{Z}}_t)$ and $I_{t-1}(\tilde{\mathbf{Z}}_{t-1})$ in consecutive frames, and its deviation from the model shape in the predicted subspace.

We define a $NW^2 \times 1$ intensity matching error vector between the current and previous frames as

$$\mathbf{\Delta} I_t = I_t(\tilde{\mathbf{Z}}_t) - I_{t-1}(\tilde{\mathbf{Z}}_{t-1}) \tag{3.14}$$

and characterize this error vector by a Gaussian distribution learned from

training data using the maximum likelihood method as

$$\mathbf{\Delta} I_t \sim \mathcal{N}(\overline{\mathbf{\Delta} I}, \mathbf{\Phi}) \tag{3.15}$$

to compute a weighted square error for intensity matching as

$$E_I = (\mathbf{\Delta} I_t - \overline{\mathbf{\Delta} I})^T \mathbf{\Phi}^{-1} (\mathbf{\Delta} I_t - \overline{\mathbf{\Delta} I}) \tag{3.16}$$

Assuming that the intensity windows centered at different feature points are independent, $\mathbf{\Phi} = diag\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N\}$, where $\boldsymbol{\phi}_k$ corresponds to the $w^2 \times w^2$ covariance matrix of the intensity window difference $\mathbf{\Delta} I_{k,t}$ at the kth feature point, and E_I reduces to

$$E_I = \sum_{k=1}^{N} (\mathbf{\Delta} I_{k,t} - \overline{\mathbf{\Delta} I_k})^T \boldsymbol{\phi}_k^{-1} (\mathbf{\Delta} I_{k,t} - \overline{\mathbf{\Delta} I_k}) \tag{3.17}$$

Simply minimizing E_I to estimate the shape can lead to unacceptable face shapes. Hence, we impose a penalty for deviation from the learned face shape model. A reasonable penalty function to encourage conformity to face shape is the Mahalanobis distance,

$$E_S = (\mathbf{Z}_t - \boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1} (\mathbf{Z}_t - \boldsymbol{\mu}_i) \tag{3.18}$$

where $\mathbf{Z}_t$ is the normalized version of $\tilde{\mathbf{Z}}_t$, $\boldsymbol{\mu}_i$ and $\mathbf{C}_i$ are the learned mean and covariance of the subspace S_t^i, predicted for time t, using the normalized track history, $\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_{t-1}$. S_t^i is used to constrain $\mathbf{Z}_t$ and is found as

the most probable subspace, given the tracking to $t-1$:

$$i = \underset{k}{\operatorname{argmax}}\, p(S_t^k|\mathbf{Z}_{0:t-1}) \tag{3.19}$$

These probabilities can be computed as

$$p(S_t^k|\mathbf{Z}_{0:t-1}) = \sum_{j=1}^{K} P(S_t^k|S_{t-1}^j)p(S_{t-1}^j|\mathbf{Z}_{0:t-1}) \tag{3.20}$$

where, assuming conditional independence, we have

$$p(S_{t-1}^j|\mathbf{Z}_{0:t-1}) = \frac{p(\mathbf{Z}_{t-1}|S_{t-1}^j)p(S_{t-1}^j|\mathbf{Z}_{0:t-2})}{p(\mathbf{Z}_{t-1}|\mathbf{Z}_{0:t-2})} \tag{3.21}$$

$$p(\mathbf{Z}_{t-1}|\mathbf{Z}_{0:t-2}) = \sum_{j=1}^{K} p(\mathbf{Z}_{t-1}|S_{t-1}^j)p(S_{t-1}^j|\mathbf{Z}_{0:t-2}) \tag{3.22}$$

Here $p(S_t^k|S_{t-1}^j)$ is obtained from the learned subspace transition probabilities and $p(\mathbf{Z}_{t-1}|S_{t-1}^j)$ from the PPCA model.

The augmented match measure for tracking can be written as:

$$E = E_I + \lambda E_S \tag{3.23}$$

where λ trades-off the relative importance of shape matching and intensity matching, and is found experimentally for best performance.

We use the iterative Gauss-Newton method to minimize E in this non-linear weighted least squares problem. Here E_I is linearized by using a first order Taylor series approximation for I_t, and the estimate at the lth

iteration is obtained as:

$$\tilde{\mathbf{Z}}_t^l = \begin{cases} \tilde{\mathbf{Z}}_{t-1}, & l = 0 \\ \tilde{\mathbf{Z}}_t^{l-1} + \tilde{\mathbf{h}}^l, & l > 0 \end{cases} \tag{3.24}$$

$$\mathbf{Z}_t^l = \mathbf{A}_t^{l-1}(\tilde{\mathbf{Z}}_t^{l-1} + \tilde{\mathbf{h}}^l) + \mathbf{b}_t^{l-1} = \tilde{\mathbf{Z}}_t^{l-1} + \mathbf{A}_t^{l-1}\tilde{\mathbf{h}}^l \tag{3.25}$$

where $\tilde{\mathbf{h}}^l$ is optimally estimated in each iteration, as described below.

Writing Eq. 3.14 as

$$\mathbf{\Delta} I_t^l = I_t(\tilde{\mathbf{Z}}_t^{l-1} + \tilde{\mathbf{h}}^l) - I_{t-1}(\tilde{\mathbf{Z}}_{t-1}) \tag{3.26}$$

and using a linear Taylor series approximation, we have

$$\mathbf{\Delta} I_t^l \simeq \mathbf{\Delta} I_t^{l-1} + \left[\frac{\partial I_t(\tilde{\mathbf{Z}}_t^{l-1})}{\partial \tilde{\mathbf{Z}}_t}\right] \tilde{\mathbf{h}}^l \tag{3.27}$$

where

$$\mathbf{\Delta} I_t^{l-1} \triangleq I_t(\tilde{\mathbf{Z}}_t^{l-1}) - I_{t-1}(\tilde{\mathbf{Z}}_{t-1}) \tag{3.28}$$

Specializing Eq. 3.27 for the window at the kth feature point, we have

$$\mathbf{\Delta} I_{k,t}^l = \mathbf{\Delta} I_{k,t}^{l-1} + \left[\frac{\partial I_t(\tilde{\mathbf{z}}_{k,t}^{l-1})}{\partial \tilde{\mathbf{z}}_{k,t}}\right] \tilde{\mathbf{h}}_k^l \tag{3.29}$$

Defining a $w^2 \times 1$ vector $\mathbf{\Delta}_{k,t}^l$ as

$$\mathbf{\Delta}_{k,t}^l \triangleq \mathbf{\Delta} I_{k,t}^{l-1} - \overline{\mathbf{\Delta} I_k} \tag{3.30}$$

and a $w^2 \times 2$ matrix $\mathbf{J}_{k,t}^l$

$$\mathbf{J}_{k,t}^l \triangleq \frac{\partial I_t(\tilde{\mathbf{z}}_{k,t}^{l-1})}{\partial \tilde{\mathbf{z}}_{k,t}} \tag{3.31}$$

we can write

$$E_I \simeq E_I' = \sum_{k=1}^{N} (\boldsymbol{\Delta}_{k,t}^l + \mathbf{J}_{k,t}^l \tilde{\mathbf{h}}_k^l)^T \boldsymbol{\phi}_k^{-1} (\boldsymbol{\Delta}_{k,t}^l + \mathbf{J}_{k,t}^l \tilde{\mathbf{h}}_k^l) \tag{3.32}$$

Writing E_S in Eq. 3.18 for the iterative algorithm as

$$E_S = (\mathbf{Z}_t^{l-1} + \mathbf{A}_t^{l-1} \tilde{\mathbf{h}}^l - \boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1} (\mathbf{Z}_t^{l-1} + \mathbf{A}_t^{l-1} \tilde{\mathbf{h}}^l - \boldsymbol{\mu}_i) \tag{3.33}$$

we can obtain $\tilde{\mathbf{h}}_k^l$ as the solution to

$$\frac{\partial E}{\partial \tilde{\mathbf{h}}_k^l} = \frac{\partial}{\partial \tilde{\mathbf{h}}_k^l} (E_I' + \lambda E_S) = 0 \tag{3.34}$$

This yields

$$\tilde{\mathbf{h}}^l = \left[\lambda \mathbf{A}_t^{l-1^T} \mathbf{C}_i^{-1} \mathbf{A}_t^{l-1} + \mathbf{J}_t^{l^T} \boldsymbol{\Phi}^{-1} \mathbf{J}_t^l \right]^{-1} \left[\lambda \mathbf{A}_t^{l-1^T} \mathbf{C}_i^{-1} (\boldsymbol{\mu}_i - \mathbf{Z}_t^{l-1}) - \mathbf{J}_t^{l^T} \boldsymbol{\Phi}^{-1} \boldsymbol{\Delta}_t^l \right] \tag{3.35}$$

The iterations are continued until $|E^l - E^{l-1}| < \tau$ with $\tau << 1$, a predefined threshold.

3.2.3 Updating of Face Shape Subspaces

To make the tracking robust with respect to face shapes of individuals not included in the training database, the PPCA model can be adapted during tracking. Rather than update the full PPCA mixture model, the procedure is simplified by assigning $\mathbf{Z}_t$ to one of the PPCA subspaces, for example, based on a maximum probability rule. Then the updated mean $\boldsymbol{\mu}_t$ and

covariance matrix $\mathbf{C}_t$ of the chosen subspace at time t can be written as:

$$k_t = k_{t-1} + 1 \tag{3.36}$$

$$\mu_t = \frac{k_{t-1}}{k_t}\boldsymbol{\mu}_{t-1} + \frac{1}{k_t}\mathbf{Z}_t \tag{3.37}$$

$$\mathbf{y}_t = \sqrt{\frac{1}{k_t}}(\mathbf{Z}_t - \boldsymbol{\mu}_{t-1}) \tag{3.38}$$

$$\mathbf{C}_t = \frac{k_{t-1}}{k_t}(\mathbf{C}_{t-1} + \mathbf{y}_t\mathbf{y}_t^T) \tag{3.39}$$

where k_t is the number of observations in the subspace, at time t, and k_0 is the number of training samples initially assigned to the subspace using Eq. 3.12.

With the updating of a subspace's mean and covariance matrix, the principal components of the subspace need to be updated also. In our case, the covariance matrices are of dimension $2N \times 2N$ (42×42) which is small enough to solve the eigenvalue problem for the principal components in $O(42^3)$ operation. However, we used the more efficient method proposed in [81] to update the subspace model in $O(q^3 + 2Nq)$ operations, where q is the subspace model order; typically q is much smaller than $2N$ to retain 95% energy.

Based on the developments in Sections 3.2.1-3.2.3, we propose and investigate two algorithms for tracking, which are summarized below.

3.2.4 Algorithm 1

In the first frame, at $t = 0$, the feature points are manually marked, and to initialize the recursions in Eqs. 3.20-3.22, we

- Assume that $P(S_t^i|\mathbf{Z}_{0:t-1}) \equiv P(S^i) = \beta_i$, the mixing weight for the

subspace S^i learned by PPCA.

- Calculate $P(\mathbf{Z}_t|\mathbf{Z}_{0:t-1})$ by obtaining $p(\mathbf{Z}_t|S_t^i)$ from the PPCA model using the normalized hand-marked features.

Hence, Algorithm 1 can be summarized as follows:

At $t = 0$,

1. Manually mark the feature points on the frame to specify $\tilde{\mathbf{Z}}_0$.

2. Compute $\mathbf{Z}_0$ and then predict the subspace S_1^i using Eq. 3.19.

   For $t > 0$,

3. Perform the iterative optimization (Section 3.2.2, Eq. 3.35) using the predicted subspace to constrain the estimate. This yields, $\tilde{\mathbf{Z}}_t$ and $\mathbf{Z}_t$.

4. Predict the subspace for the next time instant, using Eq. 3.19.

5. Update the current subspace with $\mathbf{Z}_t$ using the method described in Section 3.2.3.

6. Repeat from Step 3 until end of video sequence.

3.2.5 Algorithm 2

Here, we use the KLT algorithm [69] to obtain the raw observation of the 21 facial feature points at time t in a video sequence. The final estimate in the previous frame is used to initialize the KLT algorithm for the current frame. Without sufficient constraints, the KLT algorithm may track incorrectly in challenging situations when there is fast head motion, rapid facial feature deformations, or occlusions by hand. Hence, the final tracking result for a video frame is obtained by smoothing the KLT observations using the

best matching subspace. The latter is found through a recursive Bayesian scheme, which uses the normalized KLT observation in the current frame and the tracking history.

Here, we use $\tilde{\mathbf{Z}}_t$ and $\hat{\tilde{\mathbf{Z}}}_t$ to denote the raw KLT observation and the smoothed track, respectively, at time t. Given the KLT observation, the normalized shape $\mathbf{Z}_t$ is computed and the best matching subspace S_t^i is chosen such that

$$i = \underset{k}{\operatorname{argmax}}\, p(S_t^k|\mathbf{Z}_{0:t}) \tag{3.40}$$

These probabilities can be computed using

$$p(S_t^k|\mathbf{Z}_{0:t}) = \frac{p(\mathbf{Z}_t|S_t^k, \mathbf{Z}_{0:t-1})p(S_t^k|\mathbf{Z}_{0:t-1})}{p(\mathbf{Z}_t|\mathbf{Z}_{0:t-1})} \tag{3.41}$$

where, assuming conditional independence we have

$$P(S_t^k|\mathbf{Z}_{0:t}) = \frac{P(\mathbf{Z}_t|S_t^k)}{P(\mathbf{Z}_t|\mathbf{Z}_{0:t-1})} \sum_{j=1}^{K} P(S_{t-1}^j|\mathbf{Z}_{0:t-1})P(S_t^k|S_{t-1}^j) \tag{3.42}$$

$$P(Z_t|Z_{0:t-1}) = \sum_{k=1}^{K} P(\mathbf{Z}_t|S_t^k)P(S_t^k|\mathbf{Z}_{0:t-1}) \tag{3.43}$$

Here again, $p(S_t^k|S_{t-1}^j)$ is obtained from the learned subspace transition probabilities and $p(\mathbf{Z}_t|S_t^k)$ from the PPCA model. The normalized smoothed track $\hat{\mathbf{Z}}_t$ is obtained by projecting $\mathbf{Z}_t$ and reconstructing it from the subspace S_t^i using

$$\bar{\boldsymbol{\alpha}}^i = \mathbf{M}^{i^{-1}}\mathbf{W}^{i^T}(\mathbf{Z}_t - \boldsymbol{\mu}^i) \tag{3.44}$$

$$\hat{\mathbf{Z}}_t = \mathbf{W}^i(\mathbf{W}^{i^T}\mathbf{W}^i)^{-1}\mathbf{M}^i\bar{\boldsymbol{\alpha}}^i + \boldsymbol{\mu}^i \tag{3.45}$$

as mentioned in Section 3.2.1, and the smoothed track in image space $\hat{\tilde{Z}}$ is

then computed from $\hat{\mathbf{Z}}_t$.

Algorithm 2 can be summarized as follows:

At $t = 0$, we manually mark the feature points on the frame to specify $\hat{\tilde{\mathbf{Z}}}_0$ and normalize it to obtain $\hat{\mathbf{Z}}_0$.

For $t > 0$,

1. Use the KLT algorithm initialized with $\hat{\tilde{\mathbf{Z}}}_{t-1}$ to obtain $\tilde{\mathbf{Z}}_t$.

2. Compute normalized shape $\mathbf{Z}_t$ and predict the subspace S_t^i (using Eq. 3.40) to smooth $\mathbf{Z}_t$.

3. Obtain the normalized smoothed track $\hat{\mathbf{Z}}_t$ using Eq. 3.44 and 3.45. The final, smoothed, track $\hat{\tilde{\mathbf{Z}}}_t$ in image space is then computed from $\hat{\mathbf{Z}}_t$.

4. Update the current subspace with $\hat{\mathbf{Z}}_t$ from Step 3 using the method of Section 3.2.3.

5. Repeat from Step 1 until end of video sequence.

The major difference between Algorithm 1 and 2, is that in the former, the updated track is obtained by jointly optimizing for texture and shape matching. Whereas, in Algorithm 2, texture matching through the KLT tracker, and enforcement of the shape constraint by projection and reconstruction of the shape from the predicted subspace, take place in separate steps.

3.3 Recognition Framework

Facial expressions in ASL are described using facial feature movements and head motions [5]. The descriptions of the six expressions (shown in Fig. 3.2)

Exp.	Brow	Eye	Head
AS	Raise	Nil	Nod
NEG	Knit	Nil	Shake
RH	Raise	Widen	Tilt(left/right)
TP	Raise	Widen	Move upward
WH	Knit	Squint	Move Forward
YN	Raise	Widen	Move Forward

Table 3.1: Simplified description of the six ASL expressions (Exp.) considered: *Assertion*(AS), *Negation*(NEG), *Rhetorical*(RH), *Topic*(TP), *Wh question*(WH), and *Yes/No question*(YN). Nil denotes unspecified facial feature movements.

considered in this chapter are summarized in Table 3.1, in terms of eye, eyebrow, and head movements. Our recognition scheme uses information from these three channels to classify the facial expression, in two stages. In the first stage, the likelihoods of facial feature movements and head motions are evaluated using HMMs, and these are input to an SVM in the second stage to provide the final classification.

3.3.1 Features

Movements of the head and facial features are obtained from the tracked feature points shown in Fig. 3.1; these include both rigid and non-rigid motion. A subset of these points exhibiting rigid motion, (E_{R3}, E_{L3}), the two inner eye corners, and N_2, the bottom middle of the nose, are shown in Fig. 3.3. A reference line is defined to pass through E_{L3} and E_{R3}, and several parameters are defined as the perpendicular distances of corresponding feature points from this line. These heights/distances shown in Fig. 3.4 are:

- Seven eyebrow parameters: *Left inner brow height* (B_{IL}), *Right inner brow height* (B_{IR}), *Left middle brow height* (B_{ML}), *Right middle brow*

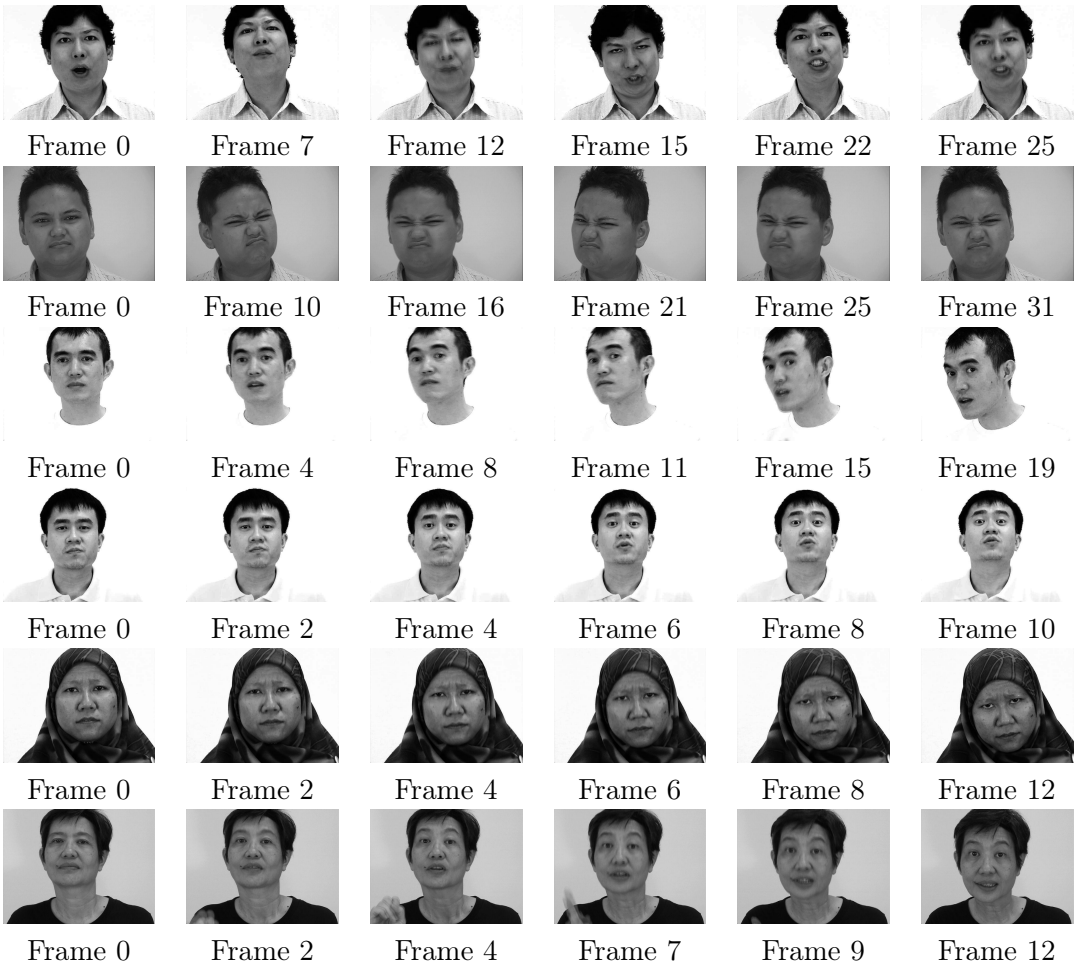


Figure 3.2: Examples of grammatical expressions. Each row shows frames from one expression. From top to bottom: AS, NEG, RH, TP, WH, YN.

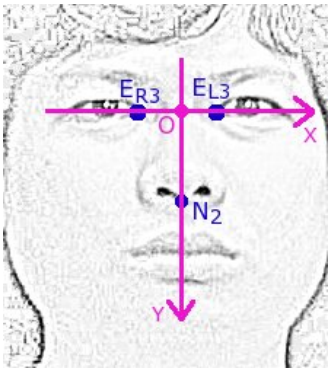


Figure 3.3: Features used for scale and in-plane rotation normalization.

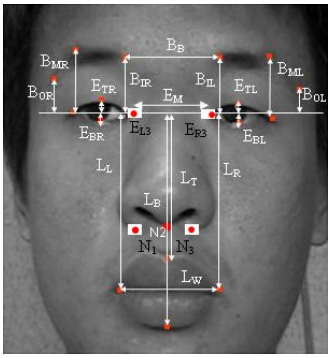


Figure 3.4: Distance features used.

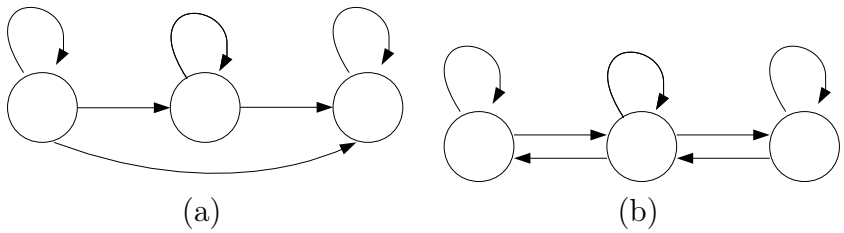


Figure 3.5: HMMs used to model facial feature movements and head motions.

height (B_{MR}), *Left outer brow height* (B_{OL}), *Right outer brow height* (B_{OR}), *Distance between brows* (B_B).

- Four eye parameters: *Left top eye height* (E_{TL}), *Right top eye height* (E_{TR}), *Left bottom eye height* (E_{BL}), *Right bottom eye height* (E_{BR}).

The features used to characterize motion of facial points are the ratios of these heights/distances and their corresponding values in the first frame. This normalization is done to remove scaling effects across video sequences.

To recognize head motions, tracks of the non-deformable facial feature locations, namely, E_{L3}, E_{R3} and N_2, are used to define three features, S_M, C_{Mx} and C_{My} as follows:

- S_M: area of the triangle formed by the above three locations in each frame.
- C_{Mx} and C_{My}: components of the 2D motion vector[1] C_M of the center of gravity of the triangle.

S_M and C_M are then normalized by E_{M0}, the distance between the two inner eye corners in the first frame, $C^n_{Mt} = \frac{C_{Mt}}{E_{M0}}$, $S^n_{Mt} = \frac{S_{Mt}}{{E_{M0}}^2}$. These 14 features obtained from the tracked facial points are used for recognition.

[1]Motion vector $\mathbf{v_{t+1}} = (x_{t+1}, y_{t+1}) - (x_t, y_t)$

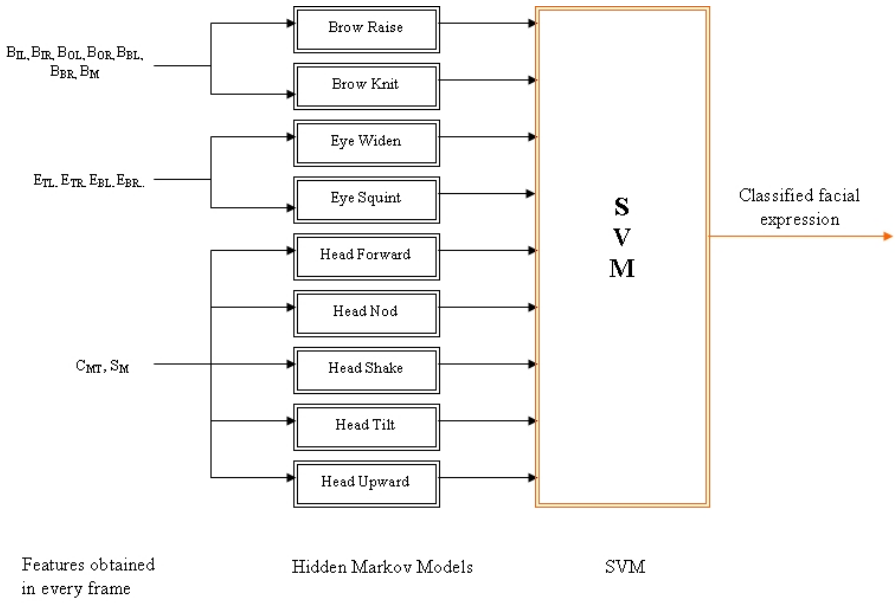


Figure 3.6: The framework for recognizing facial expressions in ASL.

3.3.2 HMM-SVM Framework for Recognition

Nine HMMs were trained for four facial feature movements (brow knit, brow raise, eye widen, eye squint) and five head motions (move forward, move upward, nod, shake, and tilt). Different HMM topologies were chosen to model the facial feature movements and head motions. We used the left-right HMM (Fig. 3.5a) to model eye and brow movements, and three head motions (move forward, move upward, tilt). Two head motions, shake and nod, were modeled by the HMM shown in Fig. 3.5b. The number of states and mixtures for each HMM were chosen experimentally using validation data. Gaussian probability density functions were used to model observations for each HMM state. Training of HMMs followed the Baum-Welch re-estimation algorithm [91].

We used two-class SVMs with Radial Basis Function (RBF) kernels to classify the facial expressions using the likelihoods of facial feature move-

ments and head motions output by the HMMs. The SVMs were trained using C-Support Vector Classification [14, 26] for every pair of classes and a voting scheme was used to obtain the final classification. The overall recognition system is illustrated in Fig. 3.6.

3.4 Experiments

We conducted several experiments to evaluate the performance of the proposed trackers and the facial expression recognition scheme.

3.4.1 Experimental Data

Videos of natural sign language facial expressions showing the signers' faces were recorded at 25 fps and spatial resolution of 640 x 480; signers were provided with appropriate signing scripts for sentences. These sentences were created or adapted from ASL resources [5, 15, 12]. Seven deaf signers from the Deaf and Hard-of-Hearing Foundation of Singapore provided the data, and each signer contributed videos in two sessions on different days. A signer signed each sentence ten times. We observed that initially some signers' facial expressions appeared forced at first, but became natural as they relaxed. The natural looking expressions were selected for our experiments. Each English sentence in the script was signed in ASL with hand signs and corresponding facial expressions, for example:

- English sentence: You know why he is crying? His mother went away!
- ASL sentence: [HE CRY]$_{TP}$ [REASON]$_{RH}$ [HIS MOTHER GO]$_{AS}$[2]

[2]subscripts denote facial expressions

Isolated facial expression sequences of the six types of grammatical markers were extracted from the video of the signed sentences, and cropped examples are shown in Fig. 3.2. The ground truth data for facial feature points and face shapes was obtained by manually marking each frame. The length of the sequences varied depending on the facial expression and the subject. The average sequence length was 18.6 frames, though there was variability between subjects and expression types. For example, the sequence length for *TP* varied between 9.4 frames to 17.9 frames, while for *AS*, the sequence length varied between 22.4 frames to 32.6 frames.

The isolated expression sequences were divided into mutually exclusive training and test sets. The training set consisted of 212 sequences, with each of the seven subjects contributing an average of five sequences for each of the six grammatical expressions. The training set was used to train all of the models used, viz: PPCA, HMMs, and SVM. Validation sets were formed from the training set to determine the configurations of HMMs and SVM. The test set consisted of 85 sequences, with an average of two sequences per subject per expression. This set was used for evaluating the performance of the trackers as well as the recognition scheme.

We also collected two other sets of video sequences especially for testing the trackers, which we refer to as the challenging set and random set. The facial features in these sequences were manually marked to create the ground truth. The challenging set contained 13 sequences of different lengths with a total of 1200 frames. These sequences exhibited complexities such as motion blur, heavy occlusions, and multiple head motions, and were obtained from ASL facial expressions we had recorded earlier, as well as from Boston University [80, 78] (Fig 3.7). The random set in-

Figure 3.7: Images from the challenging video sequences.

Figure 3.8: Images from the randomly collected video sequences.

cluded 5 sequences with 1000 frames in total. These video sequences were randomly downloaded from the Internet, and included head and shoulder shots (Fig. 3.8).

3.4.2 The PPCA Subspaces

The face shapes in the 3944 frames of the 212 training sequences were first normalized to reduce the effects of scale and in-plane rotation. Three rigid facial features were chosen for normalizing face shapes: the inner eye corners E_{R3}, E_{L3}, and the point between the two nostrils N_2 as shown in Fig. 3.3. The line passing through E_{R3} and E_{L3} was chosen to be the horizontal axis, and the line orthogonal to it and passing through N_2 is chosen to be the vertical axis, with the origin at the intersection of these axes. Using this coordinate system, the marked 2D points were translated to the origin and rotated so that the axes coincided with the image row-column axes. The feature points were then normalized by the O-N_2 distance. This is the similarity transformation of Eq.3.1.

The normalized face shapes were then partitoned into the mixture of PPCA subspaces following the method of Section 3.2.1. Since the

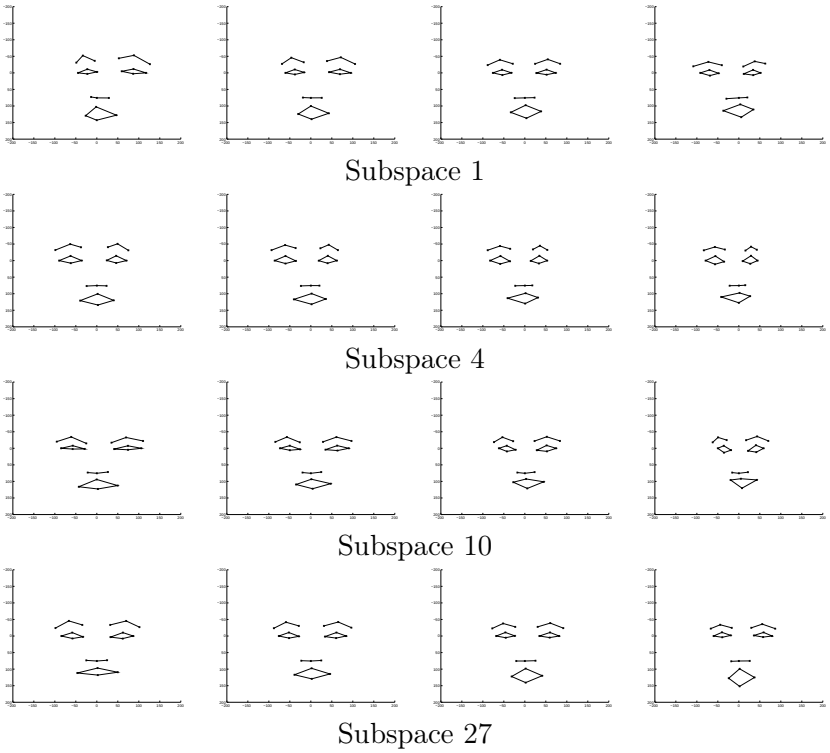


Figure 3.9: Variations of the first mode of some subspaces, showing particular deformations of face shapes due to facial feature movements and head motions that they model. Subspace 1 models a deformation of face shape when the head rotates from slightly right to slightly left and the eyebrows are knitting; Subspace 4: head rotates from frontal to left; Subspace 10: head rotates right with opening mouth and raising eyebrow; Subspace 27: head slightly rotates right with opening mouth and knitting of eyebrows.

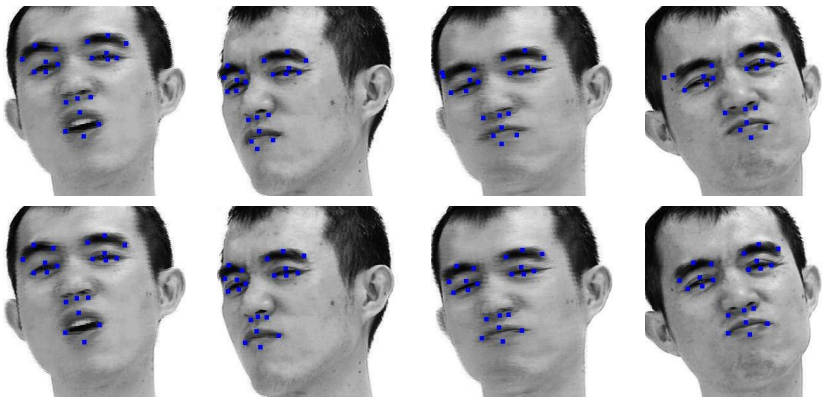
Anderson-Darling test would be applied several times in the process of predicting the number of subspaces, we chose a low significance level $\gamma = 0.0001$, with corresponding critical value of 1.8692. 28 subspaces were obtained with this setting. The principal components in the subspaces were set to retain 95% of the total energy, and this led to 13-23 components in the subspaces. The first mode of variation in a few of the subspaces is shown in Fig. 3.9. The subspace transition probabilities were obtained as described in Section 3.2.1.

3.4.3 Tracking Facial Features

In the following, we consider the tracking results from Algorithm 1, and compare them to results from KLT, Algorithm 2, etc.

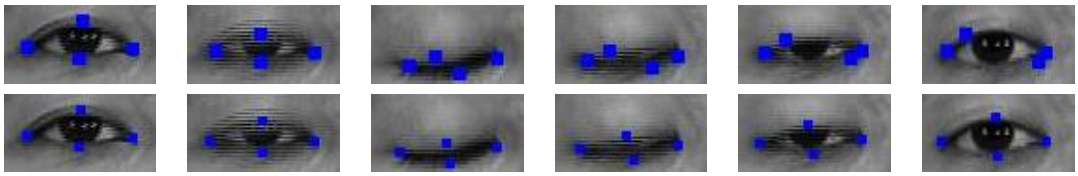
In the implementation of Algorithm 1, we found that the off-diagonal elements of the matrices ϕ^k in Eq. 3.17 were generally quite small, and hence these matrices were approximated to be diagonal. Besides, a coarse-to-fine strategy with a 3-level Gaussian pyramid was used to deal with large displacements of facial features. Algorithm 1 was employed on each level, and the tracking result obtained at the coarser level was used to initialize the algorithm at the finer level. The coarse-to-fine strategy was also applied to other tracking methods we experimented with. A suitable value of λ in Eq. 3.23 for our data set was found by experimentation, and set to be 400. Results were similar for $\lambda = 400 \pm 200$.

We first compare the performance of Algorithm 1 with the popular KLT algorithm where the latter tracks by minimizing an intensity match measure between two consecutive frames but without a shape constraint. Fig. 3.10 compares tracking by Algorithm 1 and KLT in a common ASL scenario when multiple facial feature movements and head motions occur rapidly. In Frame 10, the mouth opens and the head rotates, causing the KLT-tracked middle feature point on the lower lip to start drifting away from its true location. By Frame 15, the KLT-tracked feature points around the mouth have drifted away. It is clear that the shape constraint in Algorithm 1 results in robust tracking in this situation, even though the face in Frame 15 is far from frontal. In the rapid motion from Frame 15 to 18, the neighborhood of the right eyebrow changes rapidly, causing the KLT-tracked feature point in the middle right eyebrow to mistrack.



Frame 10 Frame 15 Frame 18 Frame 22

Figure 3.10: Tracking in an expression sequence which includes many facial feature movements and head motions. Upper row: tracking by KLT, lower row: Algorithm 1.



Frame 1 Frame 2 Frame 3 Frame 5 Frame 6 Frame 9

Figure 3.11: Algorithm 1 (lower row) can deal naturally with eye blinks due to the shape constraint, while the KLT tracks (upper row) suffer due to the rapidly changing texture in the blink region.

Narrowed eyes also cause the KLT to mistrack on the eyelids. By Frame 22, KLT-tracked feature points on the left eyebrow, eyes, nose, and mouth have drifted away. In comparison, tracking by Algorithm 1 was stable due to the shape constraint, even with rapidly changing head pose and face shape. Also, due to the shape constraint, tracks on the left eyelids which had drifted away slightly found their correct location by Frame 22.

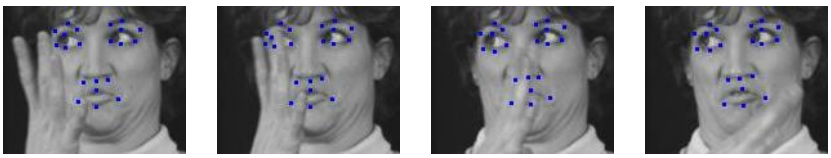
Fig. 3.11 shows that Algorithm 1 tracks points on the eye lids robustly through eye blinks due to the shape constraint, while the KLT suffers in comparison due to the rapidly changing texture in the eye area during blinks. Occlusions of the face by the hands during signing are common in

Frame 16 Frame 17 Frame 18 Frame 19 Frame 57

Figure 3.12: Algorithm 1 is stable under occlusions (lower row) while the KLT mistracks occluded points (upper row).

ASL and tracking needs to handle these situations robustly. Fig. 3.12 shows an example where the hand occludes almost half of the face during signing. The shape constraint in Algorithm 1 helps to maintain stable tracks, e.g. feature points around the right eye which are occluded by the hand are tracked; the feature point at the right mouth corner is initially affected but proceeds to its true location. Additionally, the subject in this sequence was not included in the training data. In comparison, the rapidly changing intensities in the vicinity of the tracked points due to the occluding hand cause the KLT algorithm to mistrack. Fig. 3.13 is another example of robust tracking of an unseen face occluded by hand during a different sign. In this example, feature points on the right eye and mouth are slightly affected by the occlusion but they are preserved in appropriate locations. For comparison, Fig.3.14 shows the result obtained with the original Active Appearance Model using the AAM-API library [96].

Fig. 3.15 shows tracking results by Algorithm 1 during a long sequence which includes head pose changes, facial feature deformations, eye blinks, and occlusions. After 500 frames, the feature points are still maintained at



Frame 42 Frame 43 Frame 44 Frame 45

Figure 3.13: Stable tracking by Algorithm 1 on an unseen face with occlusion by hand during signing.

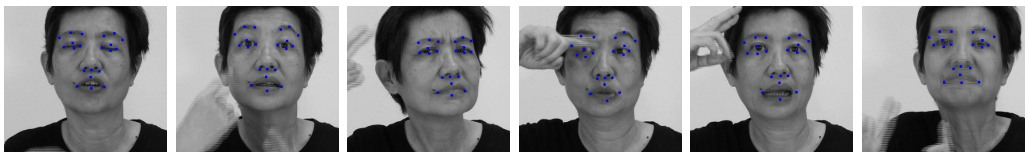
Frame 42 Frame 43 Frame 44 Frame 45

Figure 3.14: Tracking using AAM on seen face, where the AAM was trained for the person. The AAM is manually initialized on the first frame, and the result obtained in the current frame is used as the initialization for the next frame.

their appropriate locations.

Figs. 3.16-3.18 show quantitative comparisons of tracking between Algorithm 1 in Section 3.2.4, Algorithm 2 in Section 3.2.5, the KLT tracker, and Algorithm 1b - a variant of Algorithm 1 where the update procedure of Section 3.2.3 is not used and the constraining subspace is predicted by the maximum likelihood based on the latest track instead of the entire tracking history, i.e. in Eq 3.19, $i = \text{argmax}_k \, p(S_t^k | \mathbf{Z}_{t-1})$.

The tracking results were evaluated against manually labeled ground



Frame 110 Frame 140 Frame 283 Frame 353 Frame 370 Frame 500

Figure 3.15: Tracking in long sequences with multiple challenges, in order of appearance (first four images from left to right): eye blink, facial feature deformation, head rotation, occlusion.

truth for each test image frame. The average error $d_{k,t}$ of the tracking result in the tth frame of the kth video sequence was computed as $d_{k,t} = \frac{\sum_{j=1}^{j=N} |e_{k,t}^{j}|}{N}$, where $N = 21$ is the number of tracked feature points, and $e_{k,t}^{j}$ is the distance between the jth feature point and the corresponding ground truth feature point (the tracking errors are in pixel units).

Fig. 3.16 shows a comparison of the cumulative distribution of $\{d_{k,t}\}$ computed over all frames in all test video sequences, between Algorithm 1 and the other trackers listed above. Algorithm 1 is the most accurate with 90% of the displacement errors being less than 4 pixels, with Algorithm 1b (which omits the face update scheme) yielding similar performance. In comparison, Algorithm 2 which uses separate intensity matching and shape constraint steps, shows worse performance. The KLT tracker's performance was the worst, both qualitatively and quantitatively (the tracking error was less than 4 pixels with a probability of only 76%). For reference, in this data set, the face size averages about 300×300 pixels in frames of size 640×480 pixels.

The tracking stability of Algorithm 1 on the challenging data set is clearly seen in Fig. 3.17. Algorithm 1b is slightly worse than Algorithm 1, but better than Algorithm 2. The KLT tracker is the least accurate on this data set, which contains considerable motion blur and occlusion. 99.6% of the displacement errors with Algorithm 1 are within 10 pixels.

Fig. 3.18 shows the cumulative distribution of displacement errors for the video sequences randomly selected from the internet (Section 3.4.1). A displacement error of 3 pixels is obtained from Algorithm 1 in 80% of the cases, while it is 63%, 35%, and 28% in the case of Algorithm 1b, 2, and KLT, respectively. The better performance of Algorithm 1 and 1b com-

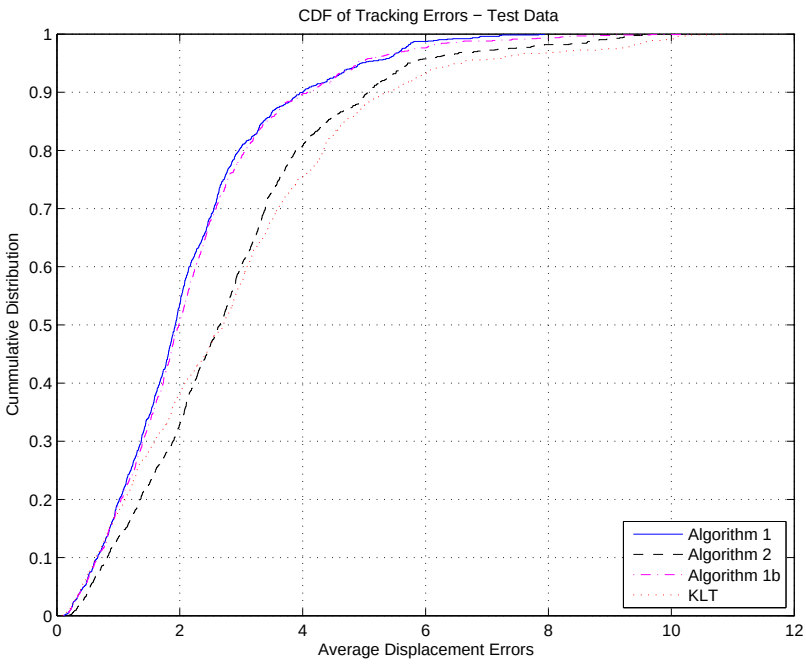


Figure 3.16: Cumulative distribution of displacement errors on the test data described in Section 3.4.1. Algorithm 1 and 1b are close in performance and better than Algorithm 2 and KLT.

pared to Algorithm 2 in all three comparisons indicates the superiority of integrated tracking with shape and texture, over separate steps for tracking and regularizing with the shape constraints. In the challenging and random data sets, which contain new faces not seen by the trackers during training, Algorithm 1b has a somewhat worse performance than Algorithm 1. This can be attributed mainly to the lack of face updates during tracking. We also obtained similar tracking results with Algorithm 1 by using the simplifying assumption $\mathbf{\Delta} I_t \sim \mathcal{N}(0, \mathbf{I})$ in Eq. 3.15. This assumption makes Algorithm 1 equivalent to integrating the KLT algorithm with shape constraints using subspaces learned by PPCA.

Algorithm 1 requires $O(1.6 \times 10^6)$ operations performed on each 640×480 frame. Using a coarse-to-fine with a 3-level Gaussian pyramid, the number of operations including those for unoptimized Gaussian filtering is

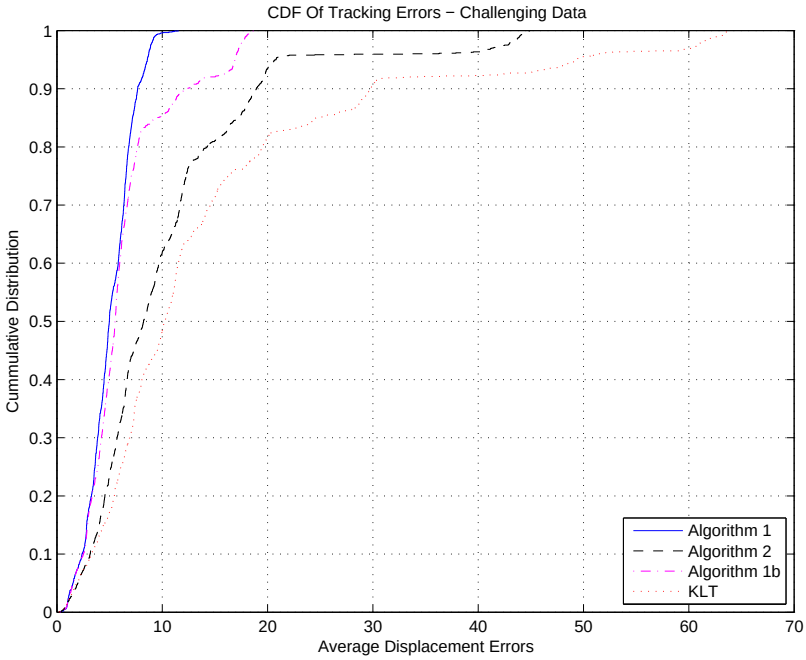


Figure 3.17: Cumulative distribution of displacement errors on the challenging data set described in Sec. 3.4.1. Algorithm 1 provides the best performance, while Algorithm 1b is slightly worse. The KLT performance is considerably worse.

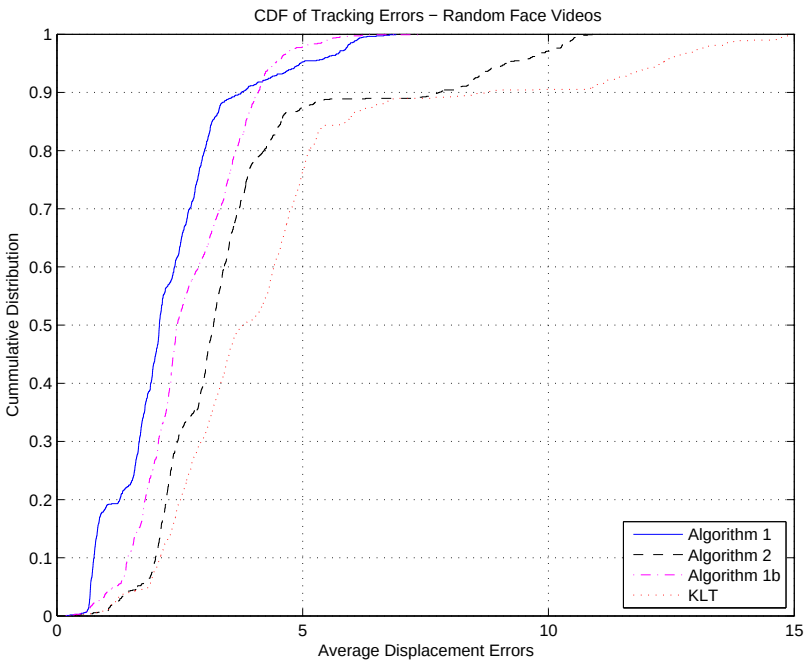


Figure 3.18: Cumulative distribution of displacement errors on the random data set (Section 3.4.1).

less than 10 millions. In another word, using state-of-the-art PC with an optimized code, our algorithm can perform in real-time.

3.4.4 Recognizing Grammatical Facial Expressions

The tracked features are used in the recognition system consisting of HMMs and the SVM described in Section 3.3.2. To obtain optimized parameters for the HMMs and the SVM, we randomly split the training data into two sets in a 75:25 ratio - a T set used for training and a V set used for validation. Parameters which provided the best result on the V set were used for obtaining results on the test data. The HMMs and SVM were trained and tested using the HMM Toolbox [75] and LIBSVM [18], respectively.

The structure of each HMM (as in Fig. 3.5) is defined by parameters M and Q, where Q is the number of states and M is the number of Gaussian mixtures of each state. We considered parameters in the range $M \in \{1, 2\}$ and $Q \in \{2, 3, 4\}$. To find the optimal structure of the nine HMMs in Fig. 3.6, we grouped the HMMs into three sets corresponding to eyebrow movement (brow raising, brow knitting), eyelid movements (eye widening, eye squinting), and head motions (move forward, move upward, nod, shake, and tilt). Inputs to each HMM were described in Section 3.3.1. We specified optimal parameters for each group of HMMs, based on the parameter set that yielded the best average accuracy for the group on the validation set. Validation sequences were classified according to the HMM that yielded the highest likelihood score. Group performance was measured by the average accuracy of all HMMs in a group over the validation set. The optimized HMMs gave accuracies of 100%, 95.74%, and 83.02% for eyebrow, eyelid,

Table 3.2: Confusion matrix for testing with MAT-MAT(%).

	YN	**WH**	**NEG**	**TP**	**AS**	**RH**
YN	78.57	0	0	14.29	0	7.14
WH	0	100	0	0	0	0
NEG	0	0	100	0	0	0
TP	0	0	0	85.71	7.14	7.14
AS	0	0	0	0	100	0
RH	7.14	0	0	0	7.14	85.71

and head movements, respectively on the validation data. The SVM was also optimized using the validation set. The inputs to the SVM are the likelihoods from the optimized HMMs. Once the optimum SVM parameters were found, the SVM was retrained using all the training data.

To assess the influence of the tracker on recognition performance we trained and tested the system with tracking inputs obtained from manually annotated tracks (MAT) and tracks obtained by using Algorithm 1 (Alg1). We used the test set of 85 isolated grammatical expression sequences described in Section 3.4.1. The average number of video sequences for each expression in this test set is 14. The recognition rates for MAT-MAT (the system trained and tested with manually annotated tracks), and Alg1-Alg1 were both 91.76%. The similarity of results for MAT-MAT and Alg1-Alg1 suggests that the tracker using Algorithm 1 can track facial feature points for facial expression classification as well as the manually annotated feature points. Tables 3.2 and 3.3 show the confusion matrices obtained by classifying the test data with MAT-MAT and Alg1-Alg1, respectively. Besides, our experiments reported in [82] show that results for Alg2-Alg2 is worse than MAT-MAT. In other words, Algorithm 1 is a better choice for tracking facial features to recognize grammatical facial expressions of interest.

For comparison, we also modeled the six grammatical facial expressions

Table 3.3: Confusion matrix for testing with Alg1-Alg1(%).

	YN	**WH**	**NEG**	**TP**	**AS**	**RH**
YN	78.57	7.14	0	7.14	0	7.14
WH	0	100	0	0	0	0
NEG	0	0	100	0	0	0
TP	7.14	0	0	92.86	0	0
AS	0	0	0	0	100	0
RH	14.29	0	0	0	7.14	78.57

Table 3.4: Confusion matrix for recognizing ASL expressions by modeling each expression with an HMM on Alg1 data(%).

	YN	**WH**	**NEG**	**TP**	**AS**	**RH**
YN	78.57	0	0	0	7.14	14.29
WH	0	71.43	7.14	0	14.29	7.14
NEG	0	0	93.33	0	6.67	0
TP	7.14	0	0	71.43	21.43	0
AS	0	0	0	7.14	85.71	7.14
RH	14.29	0	0	0	7.14	78.57

using six HMMs, with the classification determined by the HMM with the highest likelihood score. The inputs to the HMMs were 14-D feature vectors consisting of the facial feature parameters (described in Section 3.3.1) extracted from each frame. The HMM structures (the number of states and the number of Gaussian mixtures per state) were optimized using the T and V sets. An average recognition rate of 80% and 83.53% was obtained on the Alg1 and MAT data, respectively. The confusion matrix for the recognition results obtained on the Alg1 data is shown in Table 3.4.

We also conducted experiments for person independent recognition. In this experiment, the HMMs and the SVM were trained with the data from all subjects except one, and the recognition system was tested on the excluded data. Average recognition rate per person using the MAT data was 87.88% while the tracked data from Algorithm 1 yielded 87.71%. The average recognition rate per expression in both cases was also comparable.

Table 3.5: Person independent recognition results with MAT data (%) (AvgS: average per subject, AvgE: average per expression).

Subject	**YN**	**WH**	**NEG**	**TP**	**AS**	**RH**	AvgS
1	85.71	100	100	100	100	100	97.62
2	100	100	100	100	100	57.14	92.86
3	14.29	100	71.43	85.71	85.71	28.57	64.29
4	87.5	100	100	100	100	85.71	95.45
5	100	100	100	85.71	100	57.14	90.7
6	85.71	100	100	100	100	100	97.62
7	71.43	100	28.57	85.71	100	75	76.74
AvgE	77.81	100	85.71	93.88	97.96	71.94	87.88

Table 3.6: Person independent recognition results using tracks from Algorithm 1 (%).

Subject	**YN**	**WH**	**NEG**	**TP**	**AS**	**RH**	AvgS
1	85.71	100	100	100	100	100	97.62
2	100	100	100	100	100	42.86	90.48
3	28.57	100	100	85.71	71.43	42.86	71.43
4	87.5	100	85.71	100	87.5	71.43	88.69
5	85.71	100	100	100	100	71.43	92.86
6	57.14	100	100	100	100	100	92.86
7	85.71	100	57.14	100	100	37.5	80.06
AvgE	75.77	100	91.84	97.96	94.13	66.58	87.71

The confusion matrices are shown in Tables 3.5 and 3.6, respectively.

Finally, we used our tracker with Algorithm 1 and a slightly modified recognition scheme to recognize the six universal facial expressions using the CMU data set [54]. The faces here are mainly frontal with minimal head motion. Also, in contrast to the ASL grammatical facial expressions, the universal expressions contain significant information in the mouth region. Due to these different characteristics, we used a modified set of HMMs for recognition: six left-to-right HMMs (Fig. 3.5a) to model mouth movements observed in the six universal facial expressions (wide open, stretched open, smile, curved lips (mouth closed), curved upper lip (mouth open), and pursed lips), one HMM to model formation of the nasolabial furrow, two

Table 3.7: Confusion matrix for recognizing six universal expressions (%).

	Surprise	**Fear**	**Happy**	**Sad**	**Disgust**	**Angry**
Surprise	83.33	5.56	0	5.56	0	5.56
Fear	6.25	81.25	12.5	0	0	0
Happy	0	0	95.24	0	0	4.76
Sad	0	0	0	93.33	0	6.67
Disgust	0	18.18	0	18.18	54.55	9.09
Angry	0	0	0	37.5	12.5	50

HMMs to model eyelid movements (eye widen, eye squint), and two HMMs to model eyebrow movements (brow raised, brow knit). The likelihood scores obtained from these HMMs were input to an SVM to classify these six expressions. The optimal structures and parameters for this recognition system were identified following the method described at the begining of this section.

To characterize the mouth movements, we used five parameters (as shown in Fig. 3.4): *Left lip corner height* (L_L), *Right lip corner height* (L_R), *Top lip height* (L_T), *Lip width* (L_W), and *Lip height* (L_B - L_T). To characterize formation of the nasolabial furrow, we used four parameters (as in Fig 3.1 and 3.4, the reference line is the line formed by E_{L3} and E_{R3}): *Left nose corner height* (distance between N_3 and the reference line), *Right nose corner height* (distance between N_1 and the reference line), *Left brow-eye distance* (distance between B_{L3} and E_{L3}), *Right brow-eye distance* (distance between B_{R3} and E_{R3}). To characterize the eyelid and eyebrow movements, we used the same sets of parameters as described in Section 3.3.1. All distance parameters were also normalized with respect to their corresponding values in the first frame.

We used 397 video sequences from 97 subjects in total. Among these, 308 sequences were used for training (number of sequences per expression:

Surprise: 58, Fear: 48, Happy: 78, Sadness: 59, Disgust: 36, Anger: 29), and 89 sequences were used for testing (number of sequences per expression: Surprise: 18, Fear: 16, Happy: 21, Sadness: 15, Disgust: 11, Anger: 8). In the total data set used, there was about one video sequence per subject per expression, so the testing was naturally person independent. The average recognition accuracy with optimized settings was 80.9%, and the confusion matrix is shown in Table 3.7 for the test sequences.

3.5 Conclusion

We proposed algorithms for tracking facial features in sign language video, assessed their performance on the basis of their tracking accuracies and also used them in a recognition system for isolated facial expressions in ASL. The robustness of our trackers derive from shape constraints learned by a mixture PPCA model. The shape constraint is governed by a Bayesian framework which predicts or selects the subspace used to restrict the face shape deformation in each frame. In Algorithm 1, the shape constraint and an intensity matching constraint are integrated into an energy-based optimization framework to stabilize tracking. In Algorithm 2, the KLT tracks are smoothed by a reconstructed shape from a recursively predicted, best matching subspace; however, here the intensity matching and shape constraints are implemented in separate steps. The results show that our algorithms can track facial features robustly under various changes of head poses, temporary facial occlusions, and significant facial feature movements. The integrated tracking scheme of Algorithm 1 yielded the best accuracy. The proposed recognition framework utilized temporal vi-

sual cues obtained from the tracker using nine HMMs, and an SVM. The SVM inputs were the HMM likelihoods of facial feature movements and head motions for identifying six isolated grammatical facial expression in ASL. The experiments showed that the recognition results of using the tracks from Algorithm 1 were as good as from the manually annotated data, with both yielding accuracy of 91.76%. Similarly, the person independent tests yielded 87.88% and 87.7% accuracy for manually annotated tracks, and tracks from Algorithm 1, respectively. Further, on the CMU facial expression database, a slightly modified feature set and recognition scheme yielded 80.9% accuracy for the six universal expressions. Using the proposed trackers, we will address the problem of recognizing continuous facial expressions in ASL.

Chapter 4

Recognizing Continuous Grammatical Markers

4.1 Introduction

In this chapter, we consider recognizing continuous facial gestures in sign language, particularly grammatical markers in ASL. The six grammatical markers considered in this paper are summarized in Table 3.1 in terms of eye, eyebrow, and head movements. We propose to use a layered Conditional Random Field (CRF) model [61] for this purpose. The classifier includes two CRF layers, the first layer to model head motions and the second to model facial expressions. The separate head motion layer helps to reduce the ambiguity in recognizing facial expressions in the second layer. For each video sequence, probabilities of different head motions are evaluated by the first layer, and these are input to the second layer together with other features for labeling the grammatical marker in each frame. Manually annotated labels of head motions and grammatical markers were used for

training the classifier and assessing performance. The result were compared with a HMM-based classifier. The proposed classifier yielded precision and recall rates of 94.19% and 81.36%, respectively, and yielded better results than the HMM-based classifier.

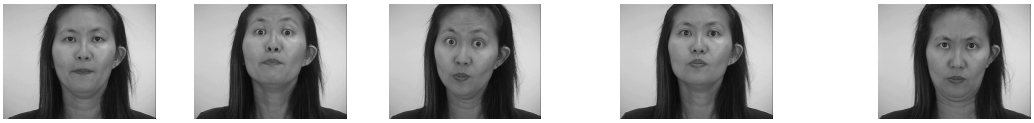
4.2 Recognizing Continuous Facial Expressions in Sign Language

4.2.1 The Challenge

Facial gestures in ASL are identified from head motion and facial feature movement. In this chapter we consider recognition of six grammatical markers listed and characterized in Table 3.1, through their gestures comprising eye, eyebrow and head movements. Here, we extend our work to recognition of continuous facial gestures as would occur in sign language discourse, and consider six types of facial gesture chains/sequences (Table 4.2) composed of these grammatical markers. Examples of these facial gesture chains are shown in Table 4.1. The figure shows obvious variations in different people performing the same grammatical markers, e.g. Yes/No question or Assertion in different chains. In this figure, unidentified expressions correspond to those not in the grammatical marker set and usually to transition expression between a pair of markers. Besides, the Neutral expression is also not interested in the current context, and is hence labeled as an unidentified expression.

There are several aspects to the continuous facial gesture recognition problem which make it challenging, more so than isolated recognition.

Table 4.1: Examples of six types of grammatical marker chains. The neutral expression shown in the first frame is not related to grammatical markers, and is considered to be an unidentified expression. An unidentified facial gesture can also be present between any two grammatical markers and can vary greatly depending on nearby grammatical markers.



Unidentified	Topic	Unidentified	Assertion	

Unidentified	Topic	Unidentified	Negation	

Unidentified	Topic	Unidentified	Yes/No question	

Unidentified	Topic	Rhetorical	Assertion	

Unidentified	Topic	Wh question	Unidentified	Yes/No question

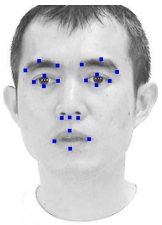
Unidentified	Topic	Unidentified	Yes/No question	Assertion

Table 4.2: Different types of grammatical marker chains considered.

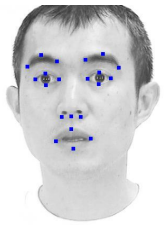
Chain	English sentence	ASL signs
TP AS	I really want the book!	[BOOK]$_{TP}$ [WANT]$_{AS}$
TP YN	Do you want the book?	[BOOK]$_{TP}$ [WANT]$_{YN}$
TP NEG	I don't want the book.	[BOOK]$_{TP}$ [WANT]$_{NEG}$
TP RH AS	I know where the game is! It's in Singapore.	[GAME]$_{TP}$ [WHERE]$_{RH}$ [SINGAPORE]$_{AS}$
TP WH YN	Where is the game? Is it in New York?	[GAME]$_{TP}$ [WHERE]$_{WH}$ [NEW YORK]$_{YN}$
TP YN AS	Do you know that book? I finished it!	[BOOK]$_{TP}$ [KNOW]$_{YN}$ [FINISH]$_{AS}$

Movement epenthesis is the extra motion required by the head (and facial features), due to physical constraints, to transit from the end of the previous gesture to the neutral state before beginning to form the next grammatical marker; this is difficult to model due to its variability. *Co-articulation* refers to the appearance of a head gesture being influenced by adjacent gestures. Speech also has the co-articulation effect, but not movement epenthesis. There can also be asynchronization between head motion and facial feature movement. Movement epenthesis and co-articulation effects between grammatical markers are shown in Tables 4.3-4.5. The example shows the grammatical marker chain *TP WH YN* when a subject is signing the words "Game", "Where", "New York" to convey the English sentences, "Where is the game? Is it in New York?".

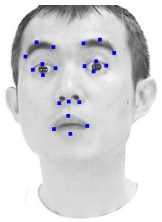
Table 4.3: A subject's facial gestures while signing the English sentence "Where is the game? Is it in New York?". Here, his facial gestures are showing the *Topic (TP)* grammatical marker while his hands are signing the word "Game".



Frame 1 *(Und)*: The video sequence starts with a neutral expression and head at neutral position.

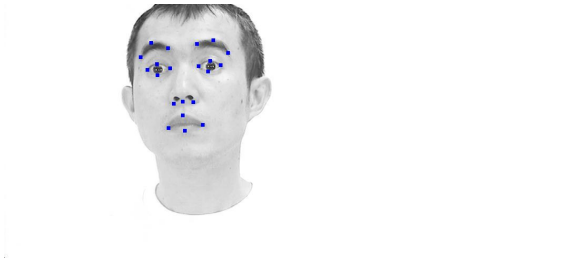


Frame 3 *(Und, the TP marker is being formed)*: The brows are being raised, the eyes are widening, while the head is still. Movement of head, brows, and eyes appear asynchronous.



Frame 7 *(TP)*: the head moves backwards together with raised brows and widened eyes.

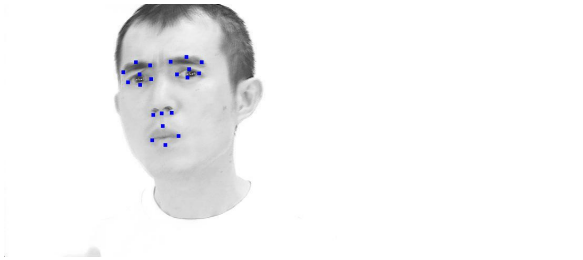
Table 4.4: (Continued from Table 4.3) The subject's facial gestures are changing from *Topic* to *Wh question (WH)* grammatical marker while his hands are signing the word "Where".



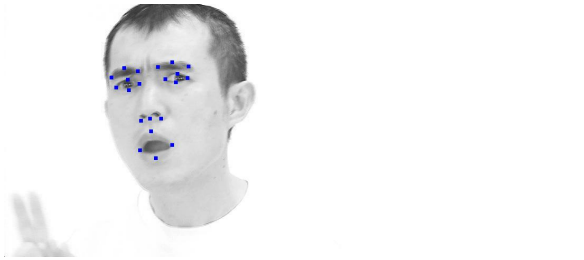
Frame 13 *(TP)*: The facial gesture is being held still because signing the word GAME has not finished.

Frame 16 *(Und)*: The head is held still while the brows and eyes are changing back to their normal states. Here, changes of eyes and brows are movement epentheses towards neutral.

Frame 19 *(Und, The Wh marker is being formed)*: The head is held still, while the brows are knitting, and the eyes are squinting.



Frame 23 *(Wh)*: Head is moving forward, slightly turning right due to the subject's habit, and past the neutral position. The head motion from Frame 19 to this frame is a movement epenthesis of the head. Besides, the *WH* expression starts when the brows have already been knit and the eyes have already been squinting (asynchronous effect).



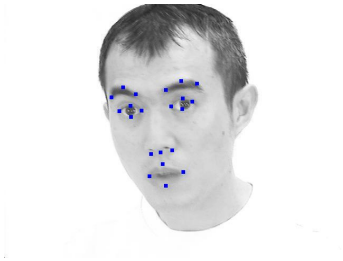
Frame 28 *(Wh)*: Head is moving forward and slightly turning right.

Frame 33 *(Wh)*: Head stops after moving forward and slightly turning right.

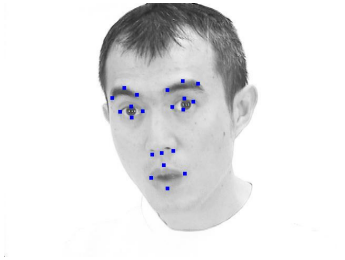
Table 4.5: Continued from Table 4.4) The subject's facial gestures are changing from *WH* to *Yes/no question (YN)* grammatical marker while his hands are signing the word "NEW YORK".

Frame 38 *(Wh)*: The *WH* marker is being held from Frame 33.

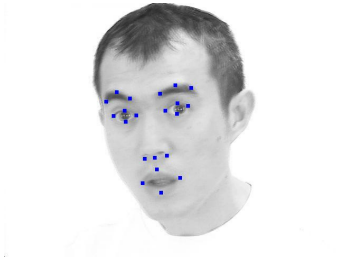
Frame 42 *(Und, the YN marker is being formed)*: The head is still while brows and eyes are being relaxed.



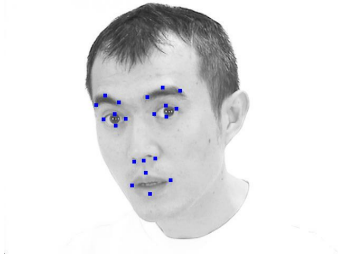
Frame 47 *(Und, the YN marker is being formed)*: Head is moving towards neutral position (movement epenthesis). Besides, the head also moves slightly downward due to the subject's habit. Eyes are widened and brows are raised.



Frame 48 *(YN)*: The head starts moving forward while brows and eyes have already been raised and widened.



Frame 51 *(YN)*: The head is moving forward slowly.

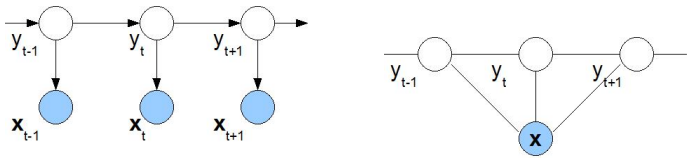


Frame 55 *(YN)*: The head stops after moving forward. This second forward head motion is not as noticeable as in the previous *WH* marker (frame 23 to 33) because the co-articulation effect from the *WH* marker causes the starting position of the head motion to not be at the comfortable, neutral, position.

Visually, the beginning and ending of an expression is considered to coincide with the beginning and ending of the head motion corresponding to that expression. However, during an expression, movements of facial features like brows and eyes are independent and may evolve asynchronously with the head motion. This asynchronization adds an uncertainty in identifying a facial expression by using a combination of features related to head motions and facial feature movements. An effective strategy to deal with this problem is to use multi-channel frameworks [84] [79], where each channel is trained to analyze a different signal, and the outputs combined to yield the final classification.

The movement epenthesis and co-articulation between grammatical markers also introduce additional types of noise. The movement epenthesis between head motions is unavoidable due to physical constraints. The head tends to move back to the neutral position to comfortably start the next motion. The asynchronization observed is in the movement of eyes and brows which tend to hold the state established at one expression into the next expression if the two expressions have similar eye and brow movements. Also, eyes and brows have to move back to their neutral positions between different states (knitting or raising eyebrows; widening or squinting eyes). Besides, the movements of the eyes and brows can be affected by factors that are not related to facial expressions of interest: natural eye blinks, hand signs for adjectives such as HUNGRY or FAST involving added facial expressions.

Furthermore, the unidentified expressions between facial expressions of interest are highly varied due to combinations of movement epenthesis and co-articulation. Thus it will be ineffective to model the expression se-



(a) HMM model (b) Linear-chain CRF model

Figure 4.1: Illustrations of HMM and linear-chain CRF models.

quences using generative models like HMMs. A discriminative model may be more suited for this scenario, and we propose to use a layered CRF model to handle head motion and facial expression.

4.2.2 Layered Conditional Random Field Model

The problem of recognizing continuous grammatical markers can be modeled as a problem of assigning a label sequence $\mathbf{y}$ composed of grammatical markers and the unidentified expression to an observation sequence $\mathbf{x}$.

This problem can be approached by using generative models like HMMs (Fig. 4.1a) which aim to maximize the joint probability $P(\mathbf{y}, \mathbf{x})$:

$$P(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^{T} P(x_t|y_t)P(y_t|y_{t-1}) \tag{4.1}$$

where T is the length of the sequence, x_t and y_t are the observation and label of frame t, respectively. This approach requires the implicit modeling of the observations, and making the assumption that observations are independent given the labels (or hidden state). If the distribution of observations is complex, the task of modeling them will add further challenge to the problem of sequence labeling.

Discriminative probabilistic models like the CRF model [61] avoid modeling the observation distribution by aiming to maximize the posterior dis-

tribution $P(\mathbf{y}|\mathbf{x})$.

The evaluation function of CRF models is composed of weighted potential functions which can utilize not only features extracted from the observations but also their interactions and temporal dependencies. In the linear-chain model (Fig. 4.1b), the probability of a label sequence $\mathbf{y}$ given an observation sequence $\mathbf{x}$ is computed as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \sum_{t=1}^{T} \left(\sum_{i=1}^{N_f} \theta_i^f f_i(y_t, \mathbf{x}) + \sum_{j=1}^{N_g} \theta_j^g g_j(y_t, y_{t-1}, \mathbf{x}) \right) \tag{4.2}$$

where f_i and g_j are potential functions that evaluate the interaction and temporal dependencies among features, respectively. N_f and N_g are the number of interaction and temporal potential functions, θ_i^f and θ_j^i are weights estimated from training data, and $Z(\mathbf{x})$ is a normalization factor given by,

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp \sum_{t=1}^{T} \left(\sum_{i=1}^{N_f} \theta_i^f f_i(y_t, \mathbf{x}) + \sum_{j=1}^{N_g} \theta_j^g g_j(y_t, y_{t-1}, \mathbf{x}) \right) \tag{4.3}$$

and can be efficiently computed using dynamic programming. A CRF can be trained by maximizing the log-likelihood of the training data set $\{\mathbf{y}^k, \mathbf{x}^k\}$

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{k=1}^{M} logP(\mathbf{y}^k|\mathbf{x}^k) && (4.4) \\ &= \sum_{k=1}^{M} \sum_{t=1}^{T} \left(\sum_{i=1}^{N_f} \theta_i^f f_i(y_t^k, \mathbf{x}^k) + \sum_{j=1}^{N_g} \theta_j^g g_j(y_t^k, y_{t-1}^k, \mathbf{x}^k) \right) && (4.5) \\ &\quad -logZ(\mathbf{x}^k) \end{aligned}$$

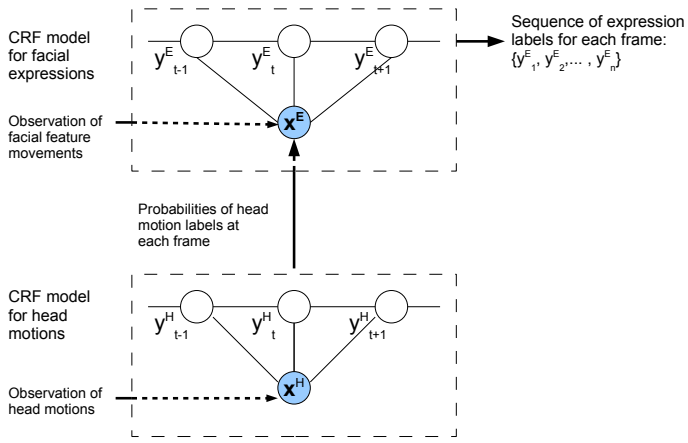


Figure 4.2: Layered CRF for recognizing continuous facial expressions in sign language.

with M is the number of sequences in the training set, and $\theta = \{\theta_i^f, \theta_j^k\}$, $i = 1, \ldots, N_f$, $j = 1, \ldots, N_g$ is the estimated parameter set of the CRF.

Lafferty et al. [61] have shown that the right hand side of Eq. 4.5 is a convex function parameterized by θ_i^f and θ_j^g, whose global optimum value can be obtained using iterative scaling algorithms [89] or gradient-based methods [73].

CRFs, which avoid the assumption of statistical independence of observations, have shown better performance than HMMs in many applications. We used a layered model of the chain CRF (Fig. 4.2) to recognize continuous facial expressions in ASL. The probabilities of head motion labels are evaluated by a CRF in the first layer. These probabilities are passed to the second layer where other facial feature channels are also integrated. The second layer CRF is trained on these integrated features, to provide expression labels for frames in the test video sequences.

Our observations show that the transition from one type of head mo-

Table 4.6: Head labels used to train the CRF at the first layer.

No.	Label	Meaning
1	Neutral (*Neu*)	Head at normal position
2	Forward (*Fw*)	Head moves forward
3	Back from Forward (*BfF*)	Head moves from forward position to neutral position
4	Backward (*Bw*)	Head moves backward
5	Back from Backward (*BfB*)	Head moves from backward position to neutral position
6	Turn left (*TL*)	Head turns left, usually a part of head shake
7	Back from Turn left (*BfTL*)	Head pose changes from leftward to frontal
8	Turn right (*TR*)	Head turns right, usually a part of head shake
9	Back from Turn right (*BfTR*)	Head pose changes from rightward to frontal
10	Move down (*MD*)	Head moves down, usually a part of head nod
11	Back from Move down (*BfMD*)	Head pose changes from downward to frontal, usually a part of head nod
12	Still	Head is kept still
13	Forward left (*FL*)	Head moves forward and slightly turns left
14	Back from Forward left (*BfFL*)	Head pose changes from leftward to frontal and head moves from forward to neutral position
15	Forward right (*FR*)	Head moves forward and slightly turns right
16	Back from Forward right (*BfFR*)	Head pose changes from rightward to frontal and head moves from forward to neutral position

tion to another can include movement epenthesis but not much articulation. Thus we choose to model movement epentheses explicitly, together with meaningful head motions. Currently, we have used 16 labels of head motions (both meaningful head motion and their movement epentheses) as described in Table 4.6 for all combinations of head motions which occur in conjunction with the six grammatical markers of interest.

In manually annotating the frames, besides the head motion label, each video frame in the data set is also labeled with one of seven facial gestures: *AS*, *NEG*, *RH*, *TP*, *RH*, *WH*, *YN*, and *UN*. The label *UN* is assigned to

frames with unidentified expressions.

As shown in Table 4.6, head motions with labels such as "Back from X" are defined to explicitly model movement epentheses. Exceptional cases are labels 7, 9, and 11 which are constituents of multi-part head motions: head shake and head nod. The *Neutral* label appears mostly at the beginning of the video sequences. During facial gestures, the head does move past the neutral position but does not stop. The frames in which the head is temporarily at the neutral position is also annotated with the *Neutral* label. The label *Still* plays an important role in segmenting meaningful head motions and their movement epentheses (Back from X) because there is usually a short pause (or even long pause) between the meaningful head motion and its "Back from" movement.

4.2.3 Observation Features

Motion of the head and facial features are obtained from the tracked feature points (shown in Fig. 3.1) using the robust tracking algorithm 1 developed in Chapter 3. The feature points are placed at both rigid and non-rigid facial locations, and distances between them are extracted and used for recognition. These distances which are similar to those in Chapter 3(see Fig. 3.4) are,

- Five eyebrow parameters: *Left inner brow height* (B_{IL}), *Right inner brow height* (B_{IR}), *Left middle brow height* (B_{ML}), *Right middle brow height* (B_{MR}), *Distance between brows* (B_B).

- Two eye parameters: *Left eye height* (summation of E_{BL} and E_{TL}), *Right eye height* (summation of E_{BR} and E_{TR}).

A reference line is defined as the line passing through the two inner eye corners, and the height parameters are the perpendicular distances of the feature points from this line. All distance parameters are normalized with respect to their corresponding values in the first frame to remove scaling effects across video sequences.

To recognize head motions, tracks of non-deformable facial feature locations, namely, the two inner eye corners (E_{L3}, E_{R3}) and the middle of the nose (N_2), are used to define three features, S_M, C_{Mx} and C_{My} as follows:

- S_M: The area of the triangle formed by the above three locations in each frame.
- C_{Mx} and C_{My}: two components of the 2D motion vector[1] C_M of the center of gravity of the triangle.

S_M and C_M are normalized by the distance E_{M0} between the two inner eye corners in the first frame $C^n_{Mt} = \frac{C_{Mt}}{E_{M0}}$ and $S^n_{Mt} = \frac{S_{Mt}}{{E_{M0}}^2}$.

These three features form the feature vector (at each frame) for the first CRF layer to evaluate probabilities of different head motions. The feature vector (at each frame) of the second CRF layer for recognizing continuous facial expressions thus has 23 elements: 16 probabilities of head motions and 7 distance ratios computed from the eyes and brows' tracked features.

4.3 Experiments and Results

Videos of natural sign language facial expressions of interest were recorded by providing signers with appropriate signing scripts for sentences. Each

[1]Motion vector $\mathbf{v_{t+1}} = (x_{t+1}, y_{t+1}) - (x_t, y_t)$

English sentence in the script was signed in ASL with hand signs and corresponding facial expressions. These sentences were created or adapted from ASL resources [5][15][12]. Deaf signers from the Deaf and Hard-of-Hearing Foundation of Singapore provided the data, and each signer contributed videos in two sessions on different days. A subject signed each sentence ten times. As mentioned in Section 4.2, our data includes six types of grammatical marker chains described in Table 4.2.

All six grammatical markers listed in Table 3.1 appear in this set of data together with 16 types of head motions described in Table 4.6. For evaluating our proposed recognition method, data from six subjects was used for experiments. The data set includes a total of 394 video sequences divided into two separate sets for training and testing. The length of each video sequence varied depending on the expression and the subject. The average number of frames in each sequence is 58.34 with standard deviation of 23.02. The longest and shortest sequences have 125 and 19 frames, respectively. Each video frame was manually transcribed to have two labels, one for the head motion, and the other for the facial expression, both identified based on observation and the signing script.

The training set consisted of 281 video sequences with an average of seven sequences from each subject for each type of expression chain. The training set was used to train both CRF layers of the model: head motion layer and grammatical marker layer. The test set consisted of the remaining 113 video sequences with an average of 3 sequences per subject per expression chain.

Recognition accuracy for facial expressions was measured by two methods: frame-based and label-aligned. In the frame-based method, the label

assigned to each frame is compared to the corresponding human annotated label. In the label-aligned method, the frame labels of each sequence are reduced such that consecutive frames with the same label are replaced by a single label. This reduced sequence of output labels is aligned to the reduced sequence of human annotated labels using the Needleman-Wunsch algorithm [77]. The number of matches, insertions, deletions, and changed labels are then obtained. Insertions are labels output by the classifier, which do not appear in the corresponding annotated data. Deletions are labels which are not recognized by the classifier while they appear in the annotated data.

The first experiment was conducted to evaluate the performance of the proposed model. The first CRF layer for head motion was trained first. The head motion probabilities output by this trained CRF was used as a part of the training vector for the CRF at the second layer. The two CRF layers were trained using the scaled conjugate gradient algorithm [73] with the CRF Toolbox [94].The output grammatical markers were obtained using Viterbi algorithm.

Figs. 4.3 and 4.4 illustrate outputs from the two CRF layers of our proposed model for the sequence shown in Tables 4.3-4.5. Fig. 4.3 shows the probability output of the first layer for the 16 head motion labels described in Table 4.6. As mentioned in Section 4.2, the head tends to move past the neutral position before starting a new motion. Fig. 4.4 shows the probability for the grammatical marker labels output by the two-layer CRF classifier. Seven probabilities including six for grammatical markers and one for the unidentified expression are obtained at each frame.

The average frame-based grammatical marker recognition rate using the

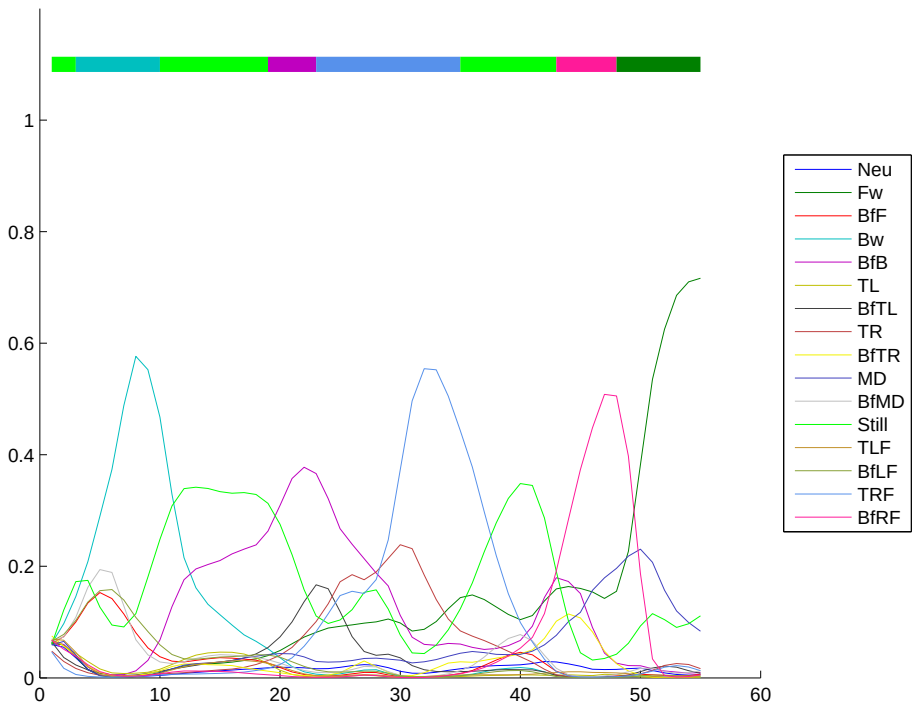


Figure 4.3: The probability outputs of the first layer CRF trained to analyze 16 types of head motion. The color bar at the top is the human annotated head motion label for this video sequence. The curve and bar with the same color are associated with the same head motion.

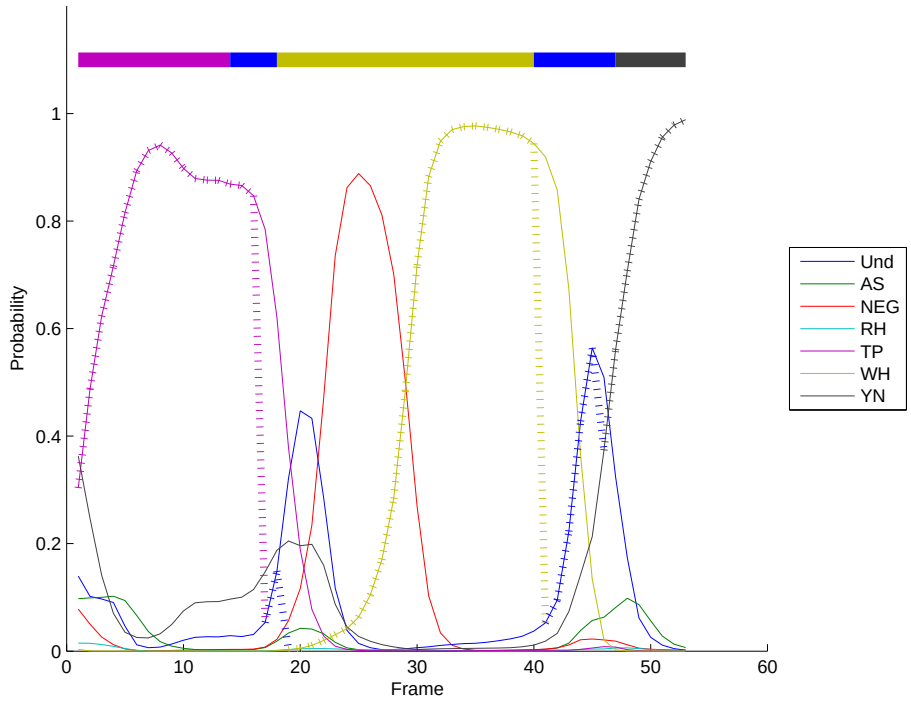


Figure 4.4: The probabilities of the grammatical markers, output by the second CRF layer trained using head motion probability output (shown in Fig. 4.3) from the first layer. The dotted curves correspond to the path chosen by the Viterbi algorithm.

complete two-layer CRF model was 76.13%. The corresponding confusion matrix is shown in Table 4.7 which shows that most of the confusions are between any grammatical marker and the unidentified expression. Particularly, frame-based label confusions usually occur at the boundary between facial gestures where ambiguous head motions and asynchronous movements of facial features are present. This makes even manual annotation of consecutive frames into different facial gestures difficult.

The label-aligned method of computing accuracy reveals more about the capability of the layered CRF for recognizing continuous grammatical markers by discounting unavoidable confusions during transitions between facial gestures. Table 4.7 can be augmented with insertion and deletion entries to obtained the extended confusion matrix $\mathbf{C}$ from which precision and recall rates are computed as follows:

- Match rate for expression i: $\mathbf{C}(i, i)$
- Change rate for expression i: $\sum_{j \notin \{i, Insert, Delete\}} \mathbf{C}(i, j)$
- Insertion rate for expression i: $\mathbf{C}(i, Insert)$
- Deletion rate for expression i: $\mathbf{C}(i, Delete)$

$$Precision = \frac{Match}{Match + Change + Insert} \tag{4.6}$$

$$Recall = \frac{Match}{Match + Change + Delete} \tag{4.7}$$

where $\mathbf{C}(i, j)$ is the value at row i and column j of the extended confusion matrix.

The extended confusion matrix for the first experiment is shown in Table. 4.8, which yields label-aligned average precision of 95.33% and average

Table 4.7: Confusion matrix obtained by labeling grammatical markers (%) with the proposed model. The average frame-based recognition rate is 76.13%.

	Und	*AS*	*NEG*	*RH*	*TP*	*WH*	*YN*
UN	65.45	6.19	1.52	4.68	14.21	2.67	5.28
AS	9.20	84.39	1.25	0.00	1.15	0.00	4.02
NEG	3.21	0.00	96.47	0.00	0.32	0.00	0.00
RH	18.62	1.79	0.00	79.34	0.26	0.00	0.00
TP	7.70	1.31	0.06	1.07	89.73	0.00	0.12
WH	22.96	0.00	0.00	4.44	0.00	69.63	2.96
YN	30.99	5.75	0.80	0.64	13.90	0.00	47.92

recall of 78.86%. Here, we can notice that there are many deletions of unidentified expressions *UN* while there are relatively few confusions between it and the grammatical markers. Because missing *UN* does not affect our primary goal of recognizing continuous grammatical markers, the performance of our proposed model can be better judged by not including the recognition results of *UN* in our final result. The *UN* labels classified as grammatical markers are considered to be insertion errors and the grammatical marker labels classified as *UN* are considered as deletion errors for the corresponding markers. From this point of view, the proposed model yields precision and recall rates of 94.19% and 81.36% respectively. The precision rate appears quite reasonable given the complexity of the problem. Besides, in this model, head motions are a strong cue for switching between facial expression. The lower recall rate hints that the layered CRF is less sensitive to change of facial gestures in video sequences. This may be improved with more descriptive features for head motion and facial feature movements.

In the second experiment, we used a single-layer CRF for recognizing continuous grammatical markers. The observation x_t at each frame com-

Table 4.8: Extended confusion matrix obtained by label-aligned grammatical marker recognition (%) using two-layer CRF model.

	UN	*AS*	*NEG*	*RH*	*TP*	*WH*	*YN*	Insert	Delete	Precision	Recall
UN	73.17	0.00	0.61	0.00	0.00	0.00	0.00	1.83	24.39	96.77	74.53
AS	0.00	83.93	1.79	0.00	0.00	0.00	8.93	0.00	5.36	88.68	83.93
NEG	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
RH	0.00	0.00	0.00	88.89	0.00	0.00	0.00	0.00	11.11	100.00	88.89
TP	0.00	0.00	0.00	0.00	91.96	0.00	0.00	0.00	8.04	100.00	91.96
WH	0.00	0.00	0.00	0.00	0.00	72.22	5.56	0.00	22.22	92.86	72.22
YN	0.00	7.02	0.00	1.75	1.75	0.00	52.63	0.00	36.84	83.33	52.63
Average										95.33	78.86

Table 4.9: Extended confusion matrix for label-aligned grammatical marker recognition result (%) using a single-layer CRF model.

	UN	*AS*	*NEG*	*RH*	*TP*	*WH*	*YN*	Insert	Delete	Precision	Recall
UN	30.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.00	100.00	30.00
AS	0.00	42.86	5.36	1.79	0.00	1.79	3.57	0.00	44.64	77.42	42.86
NEG	5.56	38.89	16.67	0.00	0.00	5.56	5.56	0.00	27.78	23.08	16.67
RH	0.00	0.00	0.00	27.78	0.00	0.00	5.56	0.00	66.67	83.33	27.78
TP	0.00	0.00	0.00	0.00	71.68	0.00	0.00	0.88	27.43	98.78	72.32
WH	0.00	0.00	0.00	0.00	0.00	44.44	0.00	0.00	55.56	100.00	44.44
YN	0.00	7.02	0.00	5.26	3.51	0.00	43.86	0.00	40.35	73.53	43.86
Average										87.72	43.57

prises the three features from head motions and seven features from eye and brows as described in Section 4.2.3. The same set of training and testing data was used for training this model and evaluating its performance. The extended confusion matrix for this experiment is shown in Table 4.9. The average precision and recall rates are 87.72% and 43.57%, respectively. Without including the recognition result for *UN*, the precision and recall rates are 84.39% and 52.33%, respectively. The significant drop of the recall rate from 81.36% in the first experiment to 52.33% suggested that head motion and facial feature movement are best analyzed as separate channels before combining them for final recognition of the grammatical markers.

In the third experiment, we applied the layered-HMM model introduced by Oliver et al. [83] [84] to our problem. The authors proposed a two-layer HMM model for recognizing human activity in an office environment. Each layer was composed of multiple HMMs, with each trained to model a single type of signal. HMMs in the first layer were trained to classify signals using features extracted from audio and video streams. Some of these classes were human speech, music, phone ringing (from audio observations), one person, multiple people, nobody (classified from video observations). Outputs from HMMs in the first layer were combined to construct observations for the second layer. At frame t, two types of outputs from HMMs in the first layer could be obtained: probabilistic or signal-based. The probabilistic output at each frame was a vector composed of probability evaluations from all HMMs at that frame. The signal-based output was a combination of two indices of the HMMs. In this type of output, HMMs in the first level were considered to be composed of two groups, audio HMMs and video HMMs. With each group, the index of the HMM yielding the maximum likelihood at frame t would be included in the signal-based output at frame t. HMMs in the second layer were trained to evaluate the presence of office activities such as "presentation", "phone conversation", or "nobody around". The output of any HMM at frame t was based on observations $\mathbf{x}^L_{t-N_L-1}, \ldots, \mathbf{x}^L_t$ which were a portion of input observations to layer L. N_L is a predefined length of sub-sequences analyzed by HMMs at layer L, N_L is experimentally defined and can be increased at higher levels which needs to analyze signals with more abstract information. In other words, each signal was analyzed using observations within a window with size N_L, and this window would be slid frame by frame towards the end of the input

observation sequence. Because each HMM was trained to analyze one single signal, by employing such slide window mechanism, each HMM can provide continuous evaluations for a sequence including different signals. However, the transition probabilities between signals were not considered.

Based on our HMM-SVM classifier introduced in the previous chapter for recognizing isolated grammatical markers, we replace the SVM in the second layer by HMMs to form a layered architecture for continuous grammatical marker recognition. We use similar HMMs for modeling head motions and facial feature movements described in the previous chapter in the first layer. The second layer consists of 4-state forward HMMs as suggested in [84]. In our case, there are 20 HMMs in the first layer to analyze the 16 types of basic head motions (as listed in Table 4.6), 2 types of eye movements (squint and widen), and 2 types of eye brow movements (knit and raised). There are 7 HMMs in the second layer for evaluating the 6 types of grammatical markers (listed in Table 3.1) and the unidentified expression. Based on our observations on the durations of facial feature movement, head motions, and grammatical markers, N_L used for the first and second layer are $N_1 = 3$ and $N_2 = 5$, respectively. The grammatical marker label for each frame is chosen based on the HMMs yielding maximum likelihood at the second layer.

Table 4.10 shows the confusion matrix for frame-based recognition results of the third experiment. As expected, the accuracy for unidentified expression is poor due to its highly variable appearance. Table 4.11 shows the extended confusion matrix using the label-aligned method; the average precision and recall rates are 38.42% and 76.39%, respectively. Without including the recognition results of *UN*, the precision and recall rates are

Table 4.10: Confusion matrix for labeling grammatical markers with the layered-HMM model. The average frame-based recognition rate is 50.05%.

	Und	*AS*	*NEG*	*RH*	*TP*	*WH*	*YN*
UN	10.77	8.51	7.71	11.14	43.15	6.30	12.42
AS	6.86	28.65	16.32	8.16	14.50	16.84	8.68
NEG	6.32	8.62	45.69	7.47	12.93	15.23	3.74
RH	4.59	3.06	5.10	65.82	13.52	3.32	4.59
TP	2.75	1.29	3.39	1.37	89.74	0.00	1.45
WH	5.68	14.32	8.89	2.47	4.20	62.96	1.48
YN	9.07	3.97	2.83	8.22	29.18	0.00	46.74

Table 4.11: Extended confusion matrix for label-based grammatical marker recognition result (%) using layered-HMM.

	UN	*AS*	*NEG*	*RH*	*TP*	*WH*	*YN*	Insert	Delete	Precision	Recall
UN	47.12	3.85	7.21	3.85	3.85	2.88	1.44	25.00	4.81	49.49	62.82
AS	0.00	23.26	1.74	1.16	0.00	2.91	2.91	67.44	0.58	23.39	71.43
NEG	0.00	0.00	17.78	0.00	0.00	1.11	1.11	80.00	0.00	17.78	88.89
RH	0.00	0.00	1.47	25.00	0.00	0.00	0.00	73.53	0.00	25.00	94.44
TP	0.00	0.00	0.00	0.00	62.07	0.00	0.00	37.36	0.57	62.43	99.08
WH	0.00	1.52	0.00	0.00	3.03	18.18	4.55	72.73	0.00	18.18	66.67
YN	0.00	2.02	0.00	4.04	6.06	0.00	39.39	42.42	6.06	41.94	68.42
Average										38.42	76.39

32.72% and 84.06%, respectively. The low precision rate may be caused by the high variance of continuous grammatical markers' appearance due to co-articulation, movement epenthesis, and asynchronization effects. Besides, the lack of constraints of the transition between signals may cause high insertion errors. A proper concatenation model of HMMs of the second layer may improve the recognition result. In a concatenation model, HMMs trained with single signals will be "connected" to form a long HMM, and the transition between different portions, original single HMMs, will be learned from training data.

Finally, we conducted person-independent recognition tests using the two-layer CRF model. In this fourth experiment, the classifier was trained

Table 4.12: Precision and recall rates (%) for person-independent recognition of grammatical markers in expression chains.

	1	2	3	4	5	6	Average
Precision	75.00	94.55	84.62	87.41	96.12	84.96	87.11
Recall	61.45	68.87	55.31	83.89	71.26	62.34	67.19

and tested six times. In each round it was trained on the data of five subjects and tested on the data of the left-out subject. The recognition results were computed using label-aligned method without including the unidentified expression. Six pairs of precision and recall rates are reported in Table 4.12 for person independent recognition. The average precision and recall rates are 87.11% and 67.19% respectively. Not surprisingly, the precision and recall rates have dropped, but still very reasonable, given the variability among subjects arising from there signing habits, etc. Having more signers for training will no doubt improve person independent results, but it would be more interesting to identify features, if possible, that are less sensitive to signer variations.

4.4 Conclusion

In this chapter, we addressed the problem of recognizing continuously signed grammatical markers in sign language video. A 2-layer CRF model was proposed for recognizing six common grammatical markers in ASL sentences. The first layer was trained for evaluating head motions and the second layer was trained for segmenting and recognizing the markers using the output from the first layer and measurements of facial feature movements. Data was collected using an experimental set up for capturing natural facial expressions composed of facial feature movements and

head motions without a forced "neutral" state between expressions. The performance of the complete 2-layer CRF model yielded precision rate of 94.19%, and recall rate of 81.36% for recognizing the six types of continuously signed grammatical markers. The person-independent test yielded 87.11% and 67.19%, respectively. The proposed classifier also outperformed two other classifiers: a linear chain CRF model and a layered-HMM classifier. These encouraging results show that the proposed 2-layer model is a viable scheme for recognizing continuous facial gestures in sign language. In the near future, we propose to enhance the robustness of the model by incorporating more descriptive features for identifying head motions. Other non-manual signals will be considered for further development of the system.

Chapter 5

Conclusion and Future Works

In this thesis, we addressed the problem of recognizing grammatical markers in ASL. In particular, we proposed algorithms for tracking facial features in sign language video, assessed their performance on video recorded from deaf signers, and used the tracked data in classification systems for isolated and continuously signed grammatical marker facial gestures in ASL.

Tracking facial features was considered for analyzing both facial feature movements and head motions which are concurrent components of grammatical markers in ASL. We first developed algorithms to track facial features robustly in the challenging sign language scenario which includes motion blur, rapidly changing head pose, occlusions, etc, which can cause frame-based intensity matching algorithms such as KLT to easily mistrack.

We used 21 feature points selected at rigid and non-rigid facial locations, and used this set of points to describe face shape. For robust tracking, the tracks must be constrained to conform to face shape. We explored two alternative algorithms for tracking. In Algorithm 1, we propagated tracks from one frame to the next by joint optimization of intensity matching

of feature points and shape constraint implemented through energy minimization. In Algorithm 2, we used the KLT algorithm to propagate the track, and then refined it using the shape constraint. Hence in the latter algorithm, the tracking and constraint enforcement are implemented in two separate steps. We used the mixture PPCA model to represent face shapes. The advantage of PPCA is that unlike other clustering and PCA-based schemes, it provides the likelihoods of face shapes belonging to particular subspaces. This is useful information that can be used in probabilistic tracking schemes. In particular, we used it to recursively predict the most probable face subspace to constrain the tracked feature points. We also used the incremental PPCA to update the mixture PPCA model to adapt to new persons, not seen during training.

We used the above two trackers and minor variants of them on facial sign language video recorded using subjects from the Deaf and Hard of Hearing Association of Singapore, while signing ASL. We also used "talking heads" video selected randomly from the Internet. The results show that our algorithms can track facial features robustly under rapid changes of head pose, temporary facial occlusions, and significant facial feature movements. The integrated tracking in Algorithm 1 yielded excellent tracking results, while Algorithm 2 was somewhat worse when the video was challenging. It is useful to note here that the tracking algorithms are generic, and can be used to handle other classes of rigid and non-rigid objects besides faces.

Our next contributions were methods for recognizing grammatical markers using the tracks of the feature points provided by Algorithm 1. We first recognized isolated grammatical markers, and then used the insights obtained to develop a classifier for continuously signed grammatical markers.

For isolated grammatical marker recognition, we used a bank of nine HMMs to separately recognize head motion and facial feature movements, using features derived from the rigid and non-rigid feature points, respectively. The likelihoods output by the HMMs were fused in an SVM classifier, trained to output the grammatical marker. The experiments showed that the recognition results of using the tracks from Algorithm 1 were as good as those from the manually annotated data, with both yielding accuracy of 91.76%. Similarly, person independent tests yielded 87.88% and 87.7% accuracy for recognition from manually annotated tracks, and tracks from Algorithm 1, respectively. Further, on the CMU facial expression database, a slightly modified feature set and recognition scheme yielded 80.9% accuracy for the six universal facial expressions.

The good recognition performance for isolated grammatical markers makes the case for a layered classifier architecture for recognition, rather than one using features from combined head motion and facial feature movements. The success of the discriminative CRF models motivated us to apply this to the problem of continuous grammatical marker recognition. Hence, we proposed a two-layer CRF model for this purpose, and compared its performance with a single layer CRF model as well as a layered HMM-based classifier.

In the two-layer CRF model, the first layer was trained for evaluating head motions and the second layer was trained for segmenting and recognizing grammatical markers using the output from the first layer and measurements of facial feature movements. The performance of the complete two-layer CRF model yielded precision rate of 94.19%, and recall rate of 81.36% for recognizing the six types of continuously signed grammatical

markers. Experimental results showed that this classifier outperformed a single layer CRF classifier and a layered HMM classifier. These encouraging results show that the proposed two-layer CRF model is a viable scheme for recognizing facial gestures in sign language.

Future Work

The robustness of our tracking algorithms under occlusions can be enhanced by explicitly detecting occluded features using more informative descriptors as SIFT [68]. Occluded features can then be recovered using a robust face alignment method [48]. The robustness of the recognition model can be enhanced by incorporating more descriptive features as those reviewed in [76].

To build a complete system for recognizing grammatical markers, we need to automatically detect facial features of interest; available methods [28, 40] could be utilized for this purpose. Furthermore, a system for fully recognizing simple signed sentences can be developed in the near future by integrating a continuous hand sign recognition framework with ours. The recognition of other non-manual signs such as conversation regulators (eye-gaze), modifiers and non-manual lexical signs (mouthing) can be developed using features obtained from the proposed tracking algorithm. Other non-manual signals will be considered for further development of the system.

Besides application in sign language, the robustness of our facial feature tracker and facial gesture classification schemes could well be used for facial gesture analysis in unstructured environments and in multimodal human

action recognition systems.

Bibliography

[1] Debra Aarons. *Aspects of the syntax of American Sign Language.* PhD thesis, Boston University, 1994.

[2] Mohamed Abdel-Mottaleb, Rama Chellappa, and Azrěel Rosenfeld. Binocular motion stereo using MAP estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 321–327, New York, NY, USA, June 1993.

[3] Keith Anderson and Peter W. McOwan. A Real-Time Automated System for the Recognition of Human Facial Expressions. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 36(1):96–105, February 2006.

[4] Benjamin J. Bahan. *Non-Manual Realization of Agreement in American Sign Language.* PhD thesis, Boston University, 1996.

[5] C. Baker and D. Cokely. *American Sign Language: A teacher's Resource Text on Grammar and Culture.* Clerc Books, Gallaudet University Press, Wasington, D.C., USA, 1980.

[6] Marian Stewart Bartlett. *Face Image Analysis by Unsupervised Learning.* Kluwer Academic Publishers, Boston, USA, 2001.

[7] Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Measuring Facial Expressions by Computer Image Analysis. *Psychophysiology*, 36:253–263, March 1999.

[8] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–573, San Diego, CA, USA, June 2005.

[9] Marian Stewart Bartlett, Javier R. Movellan, and Terrence J. Sejnowski. Face Recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464, November 2002.

[10] John N. Bassili. Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face. *Journal of Personality and Social Psychology*, 37(11):2049–2058, November 1979.

[11] Hiroaki Bessho, Yoshio Iwai, and Masahiko Yachida. Detecting Human Face and Recognizing Facial Expressions Using Potential Net. In *International Conference on Pattern Recognition*, pages 5076–5079, Barcelona, Spain, September 2000.

[12] M.J. Bienvenu and Betty Colonomos. The face of American Sign Language. Videotape, Sign Media Inc.

[13] Micheal J. Black and Yaser Yacoob. Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Mo-

tion. *International Journal of Computer Vision*, 25(1):23–48, October 1997.

[14] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *The Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, USA, July 1992. ACM Press.

[15] Byron Bridges and Melanie Metzger. *Deaf Tend Your*. Calliope Press, Silver Spring, Maryland, 1996.

[16] Andrew J. Calder, A. Mike Burton, Paul Miller, Andrew W. Young, and Shigeru Akamatsu. A Principal Component Analysis of Facial Expressions. *Vision Research*, 41(9):1169–1208, April 2001.

[17] Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):322–336, April 2000.

[18] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[19] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. Learning Partially-Observed Hidden Conditional Random Fields for Facial Expression Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 533–540, Miami, FL, USA, June 2009.

[20] Ya Chang, Changbo Hu, Rogerio Feris, and Matthew Turk. Manifold Based Analysis of Facial Expression. *Image and Vision Computing*, 24(6):605–614, June 2006.

[21] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence Chen, and Thomas Huang. Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. *Computer Vision and Image Understanding*, 91(1–2):160–187, July 2003. Special issue on Face recognition.

[22] Jeffrey F. Cohn, Adena J. Zlochower, James Lien, and Takeo Kanade. Automated Face Analysis by Feature Point Tracking Has High Concurrent Validity with Manual FACS Coding. *Psychophysiology*, 36(1):35–43, 1999.

[23] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498, Freiburg, Germany, June 1998.

[24] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Training Models of Shape from Sets of Examples. In *British Machine Vision Conference*, pages 9–18, Leeds, UK, September 1992.

[25] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[26] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, September 1995.

[27] Garrison W. Cottrell and Janet Metcalfe. EMPATH: Face, Emotion, and Gender Recognition Using Holons. In *Advances in Neural Information Processing Systems 3*, pages 564–571, Denver, CO, USA, November 1990.

[28] David Cristinacce and Timothy Cootes. Feature Detection and Tracking with Constrained Local Models. In *British Machine Vision Conference*, pages 928–838, Edinburgh, UK, September 2006.

[29] David Cristinacce and Timothy F. Cootes. A Comparison of Shape Constrained Facial Features Detectors. In *The 6th IEEE International Conference on Automatic face and gesture recognition*, pages 375–380, Seoul, Korea, May 2004.

[30] Charles Darwin. *The Expression of The Emotions in Man and Animals*. J. Murray, London, 1872.

[31] John G. Daugman. Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two Dimensional Visual Cortical Filters. *Journal of the Optical Society of America A*, 2:1160–1169, July 1985.

[32] Douglas DeCarlo and Dimitris Metaxas. Optical Flow Constraints on Deformable Models with Applications to Face Tracking. *International Journal of Computer Vision*, 38(2):99–127, July 2000.

[33] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying Facial Actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, October 1999.

[34] Fadi Dornaika and Jörgen Ahlberg. Fast and Reliable Active Appearance Model Search for 3D Face Tracking. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 34:1838–1853, August 2004.

[35] Paul Ekman. *Facial Expressions, Handbook of Cognition and Emotion*, chapter 16. John Wiley & Sons Ltd., New York, USA, 1999.

[36] Paul Ekman et al. *Facial Action Coding System.* A Human Face, 2002. chapter 1 & 2.

[37] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System.* Consulting Psychologists Press, Inc., Palo Alto, USA, 1978.

[38] Irfan Aziz Essa. *Analysis, Interpretation and Synthesis of Facial Expressions.* PhD thesis, Massachusetts Institute of Technology, 1994.

[39] Irfan Aziz Essa and Alex Pentland. Coding, Analysis, Interpretation and Recognition of Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1998.

[40] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27:545–559, April 2009.

[41] Beat Fasel and Juergen Luettin. Automatic Facial Expression Analysis: a Survey. *Pattern Recognition*, 36(1):259–275, January 2003.

[42] David J. Field. Relations between The Statistics of Natural Images and The Response Properties of Cortical Cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, December 1987.

[43] Yulia Gizatdinova and Veikko Surakka. Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):135–139, January 2006.

[44] Siome Goldenstein, Christian Vogler, and Dimitris Metaxas. Statistical Cue Integration in DAG Deformable Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):801–813, July 2003.

[45] Ralph Gross, Iain Matthews, and Simon Baker. Generic vs. Person Specific Active Appearance Models. *Image and Vision Computing*, 23(11):1080–1093, November 2005.

[46] Ralph Gross, Iain Matthews, and Simon Baker. Active Appearance Models with Occlusion. *Image and Vision Computing*, 24(6):593–604, June 2006.

[47] Haisong Gu, Yongmian Zhang, and Qiang Ji. Task Oriented Facial Behavior Recognition with Selective Sensing. *Computer Vision and Image Understanding*, 100(3):385–415, December 2005.

[48] Leon Gu and Takeo Kanade. A Generative Shape Regularization Model for Robust Face Alignment. In *European Conference on Computer Vision*, pages 413–426, Marseille, France, October 2008.

[49] Greg Hamerly and Charles Elkan. Learning the k in k-means. In *Neural Information Processing System*, pages 281–288, British Columbia, Canada, December 2003.

[50] Jesse Hoey. Hierarchical Unsupervised Learning of Facial Expression Categories. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 99–106, British Columbia, Canada, July 2001.

[51] Chung-Lin Huang and Yu-Ming Huang. Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. *Journal of Visual Communication and Image Representation*, 8(3):278–290, September 1997.

[52] William Johnson and Joram Lindenstrauss. Extension of Lipschitz Mapping into A Hilbert Space. *Comtemporary Mathematics*, 26:189–206, 1984.

[53] Rana El Kaliouby and Peter Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *IEEE Computer Vision and Pattern Recognition Workshops*, pages 154–174, Washington, DC, USA, June 2004.

[54] Takeo Kanade, Yingli Tian, and Jeffrey F. Cohn. Comprehensive Database for Facial Expression Analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, Grenoble, France, March 2000.

[55] Atul Kanaujia, Yuchi Huang, and Dimitris Metaxas. Tracking Facial Features Using Mixture of Point Distribution Models. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 492–503, Madurai, India, December 2006.

[56] Atul Kanaujia and Dimitris Metaxas. Recognizing Facial Expressions by Tracking Feature Shapes. In *International Conference on Pattern Recognition*, pages 33–38, Hong Kong, September 2006.

[57] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active Countour Models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.

[58] Rob Koenen. Mpeg-4 Project Overview. Technical Report ISO/IEC JTC1/SC29/WG11, International Organisation for Standardization, 2000.

[59] Satoshi Kumura and Masahiko Yachida. Facial Expression Recognition and Its Degree Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 295–300, San Juan, Puerto Rico, June 1997.

[60] Fernando De la Torre, Joan Campoy, Zara Ambadar, and Jeffrey F. Cohn. Temporal Segmentation of Facial Behavior. In *IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[61] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional rrandom fields: Probabilistic models for segmenting and labelling sequence data. In *International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, June 2001.

[62] Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[63] Wei-Kai Liao and Isaac Cohen. Classifying Facial Gestures in Presence of Head Motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 77–82, San Diego, CA, USA, June 2005.

[64] Scott K. Liddell and Robert E. Johnson. American Sign Language: The phonological base. *Sign Language Studies*, 64:195–278, 1989.

[65] Jenn-Jier James Lien. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*. PhD thesis, University of Pittsburgh, 1998.

[66] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of Facial Expression Extracted Automatically from Video. *Image and Vision Computing*, 24(6):615–625, June 2005.

[67] Barbara L. Loeding, Sudeep Sarkar, Ayush Parashar, and Arthur I. Karshmer. Progress in Automated Computer Recognition of Sign Language. In *International Conference on Computers for Handicapped Persons*, pages 1079–1087, Paris, France, July 2004.

[68] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[69] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *International Joint Conference on Artificial intelligence–Volume 2*, pages 674–679, British Columbia, Canada, August 1981.

[70] Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic Classification of Single Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, December 1999.

[71] Kenji Mase. Recognition of Facial Expression from Optical Flow. *IEICE Transactions*, 74(10):3474–3483, 1991.

[72] Iain Matthews and Simon Baker. Active AppearanceModels Revisited. *International Journal of Computer Vision*, 60(1):135–164, November 2004.

[73] Martin F. Møller. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks*, 6(4):525–533, 1993.

[74] Tsuyoshi Moriyama, Takeo Kanade, Jing Xiao, and Jeffrey F. Cohn. Meticulously Detailed Eye Region Model and Its Application to Analysis of Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):738–752, May 2006.

[75] Kevin Murphy. Hidden Markov Model Toolbox for Matlab. `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`.

[76] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head Pose Estimation in Computer Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–606, April 2009.

[77] Saul B. Needleman and Christian D. Wunsch. A General Method Applicable to The Search for Similarities in The Amino Acid Sequence

of Two Proteins. *Journal of Molecular Biology*, 48:443–453, March 1970.

[78] Carol Neidle. SignStream^{TM}: A Database Tool for Research on Visual-Gestural Language. *Journal of Sign Language and Linguistics*, 4(1–2):203–214, 2002.

[79] Carol Neidle, Joan Nash, Nicholas Michael, and Dimitris Metaxas. A Method for Recognition of Grammatically Significant Head Movements and Facial Expressions, Developed Through Use of a Linguistically Annotated Video Corpus. In *Language and Logic Workshop, Formal Approaches to Sign Languages, European Summer School in Logic, Language, and Information*, Bordeaux, France, July 2009.

[80] Carol Neidle, Stan Sclaroff, and Vassilis Athitsos. SignStream^{TM}: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers*, 33:311–320, August 2001.

[81] Hieu T. Nguyen and Qiang Ji. Spatio-Temporal Context for Robust Multitarget Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):52–64, January 2007.

[82] Tan Dat Nguyen and Surendra Ranganath. Tracking Facial Features under Occlusions and Recognizing Facial Expressions in Sign Language. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–7, Amsterdam, Netherlands, September 2008.

[83] Nuria Oliver, Eric Horvitz, and Ashutosh Garg. Layered Representations for Human Activity Recognition. In *IEEE International Conference on Multimodal Interfaces*, pages 3–8, Pittsburgh, PA, USA, October 2002.

[84] Nuria Oliver, Eric Horvitz, and Ashutosh Garg. Layered Representations for Learning and Inferring Office Activity from Multiple Sensory Channels. *Computer Vision and Image Understanding*, 96:163–180, November 2004.

[85] Sylvie C.W. Ong and Surendra Ranganath. Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891, June 2005.

[86] Curtis Padgett and Garrison W. Cottrell. Representing Face Images for Emotion Classification. In *Neural Information Processing Systems*, pages 894–900, Denver, Colorado, USA, December 1996.

[87] Maja Pantic and Ioannis Patras. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 36(2):433–449, April 2006.

[88] Maja Pantic and Leon J.M. Rothkrantz. Expert System for Automatic Analysis of Facial Expression. *Image and Vision Computing*, 18(11):881–905, August 2000.

[89] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:380–393, April 1997.

[90] Ariadna Quattoni, Sy Bor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1848–1852, October 2007.

[91] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[92] Sami Romdhani, Shaogang Gong, and Alexandra Psarrou. A Multi-View Nonlinear Active Shape Model Using Kernel PCA. In *British Machine Vision Conference*, pages 483–492, Nottingham, UK, September 1999.

[93] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:23–38, January 1998.

[94] Mark Schmidt and Kevin Swersky. Conditional Random Field Toolbox for Matlab. `http://www.cs.ubc.ca/~murphyk/Software/CRF/crf.html`.

[95] Eero Peter Simoncelli. Distributed Representation and Analysis of Visual Motion. Technical Report 209, Massachusetts Institute of Technology, January 1993.

[96] Mikkel B. Stegmann, Bjarne K. Ersboll, and Rasmus Larsen. FAME–A Flexible Appearance Modelling Environment. *IEEE Transactions on Medical Imaging*, 22(10):1319–1331, October 2003.

[97] Michael A. Stephens. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 69(347):730–737, September 1974.

[98] Hai Tao and Thomas Huang. Connected Vibrations: A Modal Analysis Approach for Non-Rigid Motion Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 735–740, Santa Barbara, CA, USA, June 1998.

[99] Demitri Terzopoulos and Keith Waters. Analysis of Dynamic Facial Images Using Physical and Anatomical Models. In *IEEE International Conference on Computer Vision*, pages 727–732, Osaka, Japan, December 1990.

[100] Demitri Terzopoulos and Keith Waters. Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, June 1993.

[101] Michael E. Tipping and Christopher M. Bishop. Mixtures of Probabilistic Principal Component Analysers. *Neural Computation*, 11(2):443–482, February 1999.

[102] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[103] Yan Tong and Qiang Ji. Multiview Facial Feature Tracking with a Multi-modal Probabilistic Model. In *International Conference on Pattern Recognition*, pages 307–310, Hong Kong, August 2006.

[104] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Facial Feature Tracking Using A Multi-State Hierarchical Shape Model under Varying Face Pose and Facial Expression. In *International Conference on Pattern Recognition*, pages 283–286, Hong Kong, August 2006.

[105] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Robust Facial Feature Tracking Under Varying Face Pose and Facial Expression. *Pattern Recognition*, 40(11):3195–3208, November 2007.

[106] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, Winter 1991.

[107] Paul Viola and Michael Jones. Robust Real-Time Object Detection. In *Second International Workshop on Statistical and Computational Theories of Vison*, Vancouver, Canada, July 2001.

[108] Christian Vogler and Siome Goldenstein. Facial Movement Analysis in ASL. *Journal on Universal Access in the Information Society*, 6(4):363–374, January 2008.

[109] Ulrich von Agris, Jorg Zieren, Ulrich Canzler, Britta Bauer, and Karl-Friedrich Kraiss. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4):323–362, February 2008.

[110] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade. Real-Time Combined 2D+3D Active Appearance Models. In *IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 535–542, Pittsburgh, PA, USA, June 2004.

[111] Jing Xiao, Tsuyoshi Moriyama, Takeo Kanade, and Jeffrey F. Cohn. Robust Full-Motion Recovery of Head by Dynamic Templates and Re-registration Techniques. *International Journal of Imaging Systems and Technology*, 13(1):85–94, 2003.

[112] Yaser Yacoob and Larry Davis. Recognizing Facial Expressions by Spatio-Temporal Analysis. In *International Conference on Pattern Recognition*, pages 747–749, Jerusalem, Israel, October 1994.

[113] Yaser Yacoob and Larry Davis. Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, June 1996.

[114] Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting Faces in Images: a Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:34–58, January 2002.

[115] Alan L. Yuille, Peter W. Hallinan, and David S. Cohen. Feature Extraction from Faces Using Deformable Templates. *International Journal of Computer Vision*, 8(2):99–111, August 1992.

[116] Yongmian Zhang and Qiang Ji. Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, May 2005.

[117] Zhiwei Zhu and Qiang Ji. Robust Pose Invariant Facial Feature Detection and Tracking in Real-Time. In *International Conference on Pattern Recognition*, pages 1092–1095, Hong Kong, August 2006.

List of Publications

T.D. Nguyen, K.S. Wong, S. Ranganath, and Y.V. Venkatesh. Recognize Facial Expressions in Sign Language. *IEEE SMC UK-RI 5th Chapter Conference on Advances in Cybernetic Systems*, Sheffield, United Kingdom, 2006. (oral presentation)

T.D. Nguyen, S. Ranganath. Towards recognition of facial expressions in sign language: Tracking facial features under occlusion. *15th IEEE International Conference on Image Processing*, California, USA, 2008.

T.D. Nguyen, S. Ranganath. Tracking facial features under occlusions and recognizing facial expressions in sign language. *8th IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, Netherlands, 2008.

T.D. Nguyen, S. Ranganath. Recognizing Continuous Grammatical Marker Facial Gestures in Sign Language Video. *10th Asian Conference on Computer Vision*, Queenstown, New Zealand, 2010.

T.D. Nguyen, S. Ranganath. Facial Expressions in American Sign Language: Tracking and Recognition. *Pattern Recognition*. (under revision)

T.D. Nguyen, S. Ranganath. Recognizing Continuous Gramatical Markers in American Sign Language. (in preparation)