A Computational Study of American Sign Language Nonmanuals

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy
in the Graduate School of The Ohio State University

By
Carlos Fabian Benitez-Quiroz

Graduate Program in Electrical and Computer Engineering
The Ohio State University
2015

Dissertation Committee:
Professor Aleix M. Martinez, Advisor
Professor Kevin Passino
Professor Yuan F. Zheng
American Sign Language (ASL) is a multichannel communication system that involves manual components such as hand-shape and movement, and nonmanual components as body posture, head motion and facial expressions. While significant progress has been made to understand the features defining ASL manuals, after years of research, much still needs to be done to understand its nonmanual components. ASL nonmanual linguistic research has been typically addressed by manually annotating facial events (e.g., brow raising, mouth opening, among others), and comparing the frequency of such events to find some grammatical clues about a given event in a sentence or as linguist called them construction. This tedious process is difficult to scale, especially when the number of facial events and the number of samples grow. Additionally, another major obstacle to achieve this goal is the difficulty in finding correlations between facial features and linguistic features, especially since these correlations may be temporally defined. For example, a facial feature (e.g., head moves down) occurring at the end of the movement of another facial feature (e.g., brows moves up), may specify a Hypothetical conditional, but only if this time relationship is maintained. It is however unknown for many grammatical constructions the facial features that define these dynamical facial expressions of grammar. In this work, we introduce a computational approach to efficiently carry out analysis of nonmanuals. First, a computational linguistic model of the face is defined to characterize the basic components used in ASL facial and head nonmanuals. Our results verify several components of the standard model of ASL nonmanuals and, most importantly, identify several previously unreported features and their temporal relationship. Notably, our results uncovered a complex interaction between head position and mouth shape. These findings define some temporal structures of ASL nonmanuals not previously identified by other approaches.
Second, we study the hypothesis that facial expressions of negative moral judgment (i.e., contempt, anger and disgust) have evolved into a facial expression of negation regularly used as a grammatical marker in human language and that this nonverbal signal is used as a co-articulator in speech and that, in American Sign Language, it has been grammaticalized as a nonmanual marker. Our results show commonalities in the muscle activation in three different spoken languages (i.e., English, Chinese and Spanish) and across modalities (i.e., ASL) Third, we create computational tools that allows to extract facial events by detecting the facial components (e.g., eyebrows, eyelids, mouth, nose and jawline) over potentially several video sequences of signed sentences. Our results achieve state-of-the-art precision in detecting facial landmarks and it can be easily extend to other objects such as medical images. Furthermore, we modeled the movement of the facial components as continuous functions that are classified using a maximum margin functional approach. Our new algorithm can be applied to a variety of data and application such as dynamical facial expressions and body gestures. We will show that the proposed methodologies seem to be tailor for ASL applications, they can be applied to different areas such as facial expression of emotion, face shape detection, object detection and data mining among many others.
Dedicado a:

*Mis padres Ciro y Alma*

*Mis hermanos Javier y Ciro*

*Mis sobrinos Javier, Juan Sebas, y Juanes*

*A los Ponce*


*Y especialmente a mi esposa Katie (a.k.a. la mona) y mi nena que aun no tiene nombre.*
Acknowledgements

I would like to thank my advisor Aleix Martinez for teaching me all the things that I now know about science, for the talks about research in other scientific topics and for the non-academic talks. I really learned a great deal during these years and I thank him for many other responsibilities that made me a more complete scientist.

I would like to thank Dr. Yuan Zheng, Dr. Kevin Passino and Dr. Yoonkyung Lee for being part of my committee and for the feedback provided in the many stages of my PhD. I hope that you enjoy reading my manuscript as much as I enjoy researching here at Ohio State University (sorry The Ohio State University).

I also want to acknowledge the co-authors in my papers, Paulinho Gotardo, Di You, Sam Rivera, Ronnie Wilbur and Kadir Gokgoz, it was a lot of fun to write those papers with you guys.

I would like to thank the following current and former CBCSLers which I shared many scientific discussions: Onur Hamsici, Liya Ding, Shichuan Du, Yong Tao, Riuqi Zhao, Ram and Pam Pallett.

I would like to thank other ECE grad students who discussed many optimization problems with me: Felipe, Viteri, Daniel Ospina and Ana Maria.

Finally, I would like to thank the staff at the ECE department, they became kind of my family at OSU (arguments included).
Vita

1981 .......................... Born - Cali, Colombia
2004 ............... B.S. Electronics - Pontificia Universidad
                  Javeriana - Cali, Colombia
2007 ............... M.S. Electrical Engineering - University of
                  Puerto Rico - Mayaguez, Puerto Rico
2008 ............... Graduate Research Assistant Electrical
                  and Computer Engineering - The Ohio
                  State University - Columbus, Ohio

Publications


Fields of Study

Major field: Electrical and Computer Engineering
# Table of Contents

Abstract ........................................................................................................ ii
Dedication ........................................................................................................ iv
Acknowledgements .......................................................................................... v
Vita ..................................................................................................................... vi
List of Figures .................................................................................................... ix
List of Tables ...................................................................................................... xiv

1 Introduction .................................................................................................... 1
   1.1 Literature Review .................................................................................. 5

2 Discriminant Features and Temporal Structure of Nonmanuals in American Sign Language ................................................................. 8
   2.1 Summary ................................................................................................ 8
   2.2 Introduction ............................................................................................ 9
   2.3 Methodology .......................................................................................... 12
      2.3.1 Database ....................................................................................... 13
      2.3.2 Manual Annotations ....................................................................... 14
      2.3.3 Temporal Logic Description .......................................................... 17
      2.3.4 Discriminant Analysis ................................................................... 21
   2.4 Results .................................................................................................... 23
      2.4.1 Experiment 1: Constructions Discriminant Features .................. 23
      2.4.2 Experiment 2: Polarity Discriminant Features .............................. 32
   2.5 Discussion ............................................................................................... 41
   2.6 Sentences Signed in our Database ........................................................ 44

3 The Not Face: A Grammaticalization of Facial Expressions of Emotion .......... 50
   3.1 Summary ................................................................................................ 50
   3.2 Introduction ............................................................................................ 51
   3.3 Results ..................................................................................................... 53
      3.3.1 Experiment 1. Nonverbal Expression of Negation ....................... 54
      3.3.2 Experiment 2. Co-articulation of the Not Face in Speech .......... 57
      3.3.3 Experiment 3. Grammatical (nonmanual) Marker in Signing ....... 58
      3.3.4 Experiment 4. Frequency of Production ....................................... 60
List of Figures

2.1 ELAN is a computer software that allows users to view synchronized videos simultaneously and frame by frame (top of figure), facilitating manual annotations (bottom half). Our manual annotations specify where the sentence starts and ends, where each word (or concept) starts and ends, plus the shape and configural features used to uncover the linguistic model (see text).

2.2 Samples of a video sequence of a native ASL signer signing “If #Sarah have a party tomorrow”

2.3 The configural and shape positions used to define each of the nonmanuals in sentences of ASL. In 2.3(a) we show a neutral face. A neutral face is defined as one without expression where all facial muscles are relaxed (except for the eyelids which are open). 2.3(b) We consider two configural positions for the eyebrows (up and down). 2.3(c) Blinks are marked by closing the eyelids. 2.3(d) The mouth can be open or closed. 2.3(e) We also annotate mouth shape where appropriate (flat, round and other). 2.3(f) When there is teeth showing, we consider three distinct positions – closed (top and bottom teeth touching), open (not touching), touching lips (where the top teeth are over the lower lip or the bottom teeth touch the upper lip). 2.3(g)-2.3(i) We also consider the three possible rotations of the head – turns, tilts and forward/backward movements.

2.4 Visual representation of the propositions used in our coding.
2.5 A diagram describing facial feature movements/positions and the gloss for the sentence “#BRAD-I-Xi COOK FISH ON GRILL I-Xi.” For clarity, here, we have listed the events and their time intervals in order of occurrence. The top row specifies the first event, with subsequent rows listing later occurring events. The bottom row summarizes the time interval of each signed concept. This visualization facilitates the coding of the events using the propositions in $\mathcal{P}$. For example, for the figure above, it is easy to see that head moves up occurs during event brows move up, which can be compactly expressed as $\text{during}(\text{head moves up}, \text{brows move up})$. . . . . . . . . . . . . . . . . . . . . . 20

2.6 Visual representation of the ATL feature vector $\mathbf{x}$. The dark blue color indicates a low number of occurrences for an event, while a dark orange color indicates a high number of repetitions. This figure is the histogram corresponding to the example in Fig. 2.5. The feature vector entries ($x_k$) are read from left to right ($k = 1, \ldots, 49$). . . . . . . . . . . . . . . . . . . . . . 21

2.7 Mouth positions for the sentence “WHO COOK FISH ON #GRILL.” . . . 27

2.8 Computational model of polarity in Hypothetical conditionals. . . . . . . 35

2.9 Computational model of positive versus negative polarity in Wh-questions. 38

3.1 Example images of facial expressions used to convey negative moral judgment: 3.1(a) anger, 3.1(b) disgust and 3.1(c) Contempt. . . . . . . . . . . . . . . 52

3.2 A few examples of people conveying negation with a facial expression. People from different cultures use AUs 4, 17 and either 14 or 24 or both. . . . . . . 52

3.3 We test the hypothesis that compounding the expressions of negative moral judgment (anger, disgust, contempt) yields a grammaticalized universal facial expression of negation. Shown on top are the three facial expressions of anger, disgust and contempt and their corresponding AUs. The AUs with an outer box are the ones that are combined to express negation (bottom image). . . 55
3.4 Samples of spontaneous expressions of negation. 3.4(a)-3.4(b). Pictures of professional soccer players after missing an open shot. 3.4(c). A coach disagreeing with the referee. ................................................................. 56

3.5 Sample frames of the “not face” in videos of American television. ................. 59

4.1 Can you find the 10 faces in (a)? These faces are difficult to see until a face feature is detected (e.g., a nose); then the entire face becomes salient. (b) The output of a standard face landmark detector is typically restricted to a few salient points. (c) Our novel method provides dense detections that include both salient and non-salient landmarks. ................................................................. 73

4.2 Illustration of proposed methodology: The ‘Graph Model’ block shows the graph learning phase of the method (Section 3), where we model the relationship between salient and non-salient landmarks as a probabilistic graph. Thicker edges represent larger weights between fiducials. Only a subset of edges of the fully connected graph are shown. The ‘Test Image’ block highlights cases of misdetections and multiple detections for a particular fiducial. The ‘Shape Detection’ block (section 4) shows that the graph model is used along with the local feature detections to determine the most probable shape configuration. ................................................................. 74

4.3 We show example results of the derived approach. From top to bottom, the rows correspond to the shape detections for the AR database [101], the LFW database [78], and the XM2VTS database [105]. The database contains face images in unconstrained environments. Results show robustness to occlusions, pose, and lighting. ................................................................. 83
4.4 Error histograms (Euclidean distance, in pixels) for a total of 50 detected fiducial points versus the ground truth of the testing sets. The ground truth positions were obtained by manual annotation. PGA denotes our new probabilistic graph algorithm, while AAM denotes the classical AAM.

4.5 Comparison of our algorithm with AAM [34] and the local detector of [48]. Our method provides more precise detections than the AAM, and many more fiducial points than the local feature detector.

4.6 Hand Experiments on [139]. (a) Detected landmarks delineating the shape of the Hand. (b) Error histogram (Euclidean distances, in pixels) for 52 detected landmarks versus the ground truth of the testing images.

4.7 We show a denser set of fiducial positions which were inferred using the detections from Fig. 4.3 and an expanded probabilistic graph as in Section 4.2.5.

4.8 Shown here are the two functions $x_i(t)$, $i = 1, 2$, defining the spatio-temporal movement of two fiducial points in a video sequence of an American Sign Language sentence. The feature vector of the sample video is the vector of functions $F = [x_1(t), x_2(t)]^T$. In this model, our goal is to find the linear differential equations $L(F(t))$ resulting in the representation with largest possible margin between samples of different classes.

4.9 Classification accuracy of synthetic data for different percentage of samples with missing data. As the number of samples with missing data increases, the classification accuracy for the proposed approach remains nearly unchanged. In comparison, the classification accuracy of a linear SVM drops quite fast.

4.10 Classification accuracy on synthetic data for different levels of noise. Results for the proposed approach and a classical SVM classifier.
4.11 Sample frames of the datasets used in our experiments. (a) The function defining the 2D movement of the landmarks and the rotations of the head in the ASL database of [16]. (b) The movement of the landmarks in the 3D Dynamic Facial Expression Database of [17]. (c) Gestures in the Kinect dataset of [54].
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Features of the model and their entry sets.</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>Temporal relations for the events in Fig. 2.5. It is important to note in the table below that the temporal information is encoded in the description of the ATL using the propositions in $\mathcal{P}$. The temporal information is hence intrinsically coded in this table.</td>
<td>19</td>
</tr>
<tr>
<td>2.3</td>
<td>Discriminant features for Hypothetical conditionals.</td>
<td>25</td>
</tr>
<tr>
<td>2.4</td>
<td>Discriminant features for Wh-questions.</td>
<td>28</td>
</tr>
<tr>
<td>2.5</td>
<td>Discriminant features for Wh-questions postposed.</td>
<td>29</td>
</tr>
<tr>
<td>2.6</td>
<td>Discriminant features for Yes/no-questions.</td>
<td>31</td>
</tr>
<tr>
<td>2.7</td>
<td>Discriminant features for Assertions.</td>
<td>32</td>
</tr>
<tr>
<td>2.8</td>
<td>Discriminant features for polarity in Hypothetical conditionals.</td>
<td>33</td>
</tr>
<tr>
<td>2.9</td>
<td>Discriminant features for polarity in Wh-questions.</td>
<td>37</td>
</tr>
<tr>
<td>2.10</td>
<td>Discriminant features for polarity in Wh-questions postposed.</td>
<td>39</td>
</tr>
<tr>
<td>2.11</td>
<td>Discriminant features for polarity in Assertions.</td>
<td>40</td>
</tr>
<tr>
<td>2.12</td>
<td>First set of sentences. The first column shows the sentence signed by a subset of the signers as described in Table 2.16. The second column indicates the polarity of the sentence i.e., positive or negative. The last column indicates the type of sentence i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.</td>
<td>45</td>
</tr>
</tbody>
</table>
2.13 Second set of sentences. The first column shows the sentences signed by a subset of the signers as described in Table 2.16. The second column indicates the polarity of the sentence *i.e.*, positive or negative. The last column indicates the type of sentence *i.e.*, Yes/no questions, Wh-questions, Wh-questions postposed and Assertions. ................................. 46

2.14 First set of complex sentences. The first column shows the sentence signed by a subset of the subjects as described in Table 2.16. The second column indicates the polarity of the Hypothetical conditional clause and the polarity of the dependent clause *i.e.*, Yes/no questions, Wh-questions, Wh-questions postposed and Assertions. ................................. 47

2.15 Second set of complex sentences. The first column shows the sentence signed by a subset of the subjects as described in Table 2.16. The second column indicates the polarity of the Hypothetical conditional clause and the polarity of the dependent clause *i.e.*, Yes/no questions, Wh-questions, Wh-questions postposed and Assertions. ................................. 48

2.16 Sets of sentences signed by the chosen fifteen participants. The first column is the participant code. Every session was divided between two and four blocks. In each block the participant signed sentences belonging to the ones described in Tables 2.12-2.15. ................................. 49

3.1 Consistently used Action Units (AUs) by subjects of different cultures when producing a facial expression of negation. ................................. 54

3.2 Usage of the “not face” in different languages. ................................. 58

3.3 First few sentences used in Experiments 2 & 3. The first column shows the sentence signed / uttered by participants. The second column indicates the polarity of the sentence (positive or negative). The last column indicates the type of sentence (yes/no questions, wh-questions, and assertions). ................................. 69
3.4 Questions used in Experiment 2.

3.5 Additional sentences used in Experiments 2 & 3. The first column shows the sentence uttered/signed by subjects. The second column indicates the polarity of the Hypothetical conditional clause and the polarity of the dependent clause *i.e.*, Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.
Chapter 1

Introduction

The study of grammar in sign languages is of fundamental importance in many areas such as linguistics, cognitive science, education and engineering. Sign languages provide a window for the study of what formal, highly abstract and minimally required properties constitute human linguistic knowledge [30, 52, 19] e.g., what is it about the human language system that makes it surface freely and in a full-fledged manner in the manual-visual modality when input from the oral-aural modality is not available [135]. Similarly, understanding how sign languages encode grammatical rules, which are thought to be rooted in the overall human cognitive capacity but which until recently were formally defined based mostly on spoken languages, allows researchers to generalize discoveries in the cognitive sciences [106]. Additionally, the teaching of sign languages will be much facilitated once we know more about how the grammar is encoded in its manual and nonmanual components in sign production at the clausal level. In sign language research, nonmanuals refer to linguistically-controlled uses of the face, head, and body other than the hands (see [124] for a recent review).

The sign language literature has made it clear that although affective and linguistic expressions may co-occur, they are nonetheless easily distinguished by their articulation onsets and offsets with respect to the signs made with the hands, with linguistic expressions tightly coordinated with the syntactic constituents to which they relate [6, 35, 36, 111, 124, 148, 157, 158, 160, 161, 163, 166, 167]. Similarly, there are clear distinctions between the nonmanual expressions and positions used by signers as compared to those employed by sign-naive hearing people in conjunction with speaking e.g., [167].
It has been difficult to determine which facial expressions are associated with specific grammatical functions due to the fact that any given articulation could have meaning by itself or could enter into combination with other articulations to provide an unrelated meaning. The reason for this is related to the number of articulators (e.g., head, brows, eye lids, eye gaze, nose, mouth, cheeks, chin, shoulders), the options available to each (for example, the head can turn left/right, nod up/down, or tilt to the left/right side), and the multiple combinations in which they interact. Thus, sorting through all the possibilities and testing each for what may be subtle differences in meaning is a complex problem with many variables. While it is well known how the handshape, hand movement and palm orientation form the fundamental building blocks of the manual component of the sign [140, 141, 14, 23, 134], it is still unclear how head movements and facial configurations are structured and used in sign languages. Some progress has been made describing the nonmanual contribution based on, mostly, but not exclusively [158, 99, 111, 124], painstaking and slow annotation tools [163, 162, 148, 2, 31, 36, 124, 135, 157], but there is still much to be discovered about nonmanuals, especially with the help of more efficient research tools and procedures that are instrumentally-based and ideally automatic [81, 96, 40, 125, 62]. The development of computational approaches that can assist with this process will be of great benefit to linguistic analysis.

To better understand the need for computational tools for the linguistic analysis of nonmanuals, let us review their use in sign languages. The nonmanuals used in sign languages serve a variety of functions similar to those performed by intonation or word order changes in a spoken language like English. For example, to make a question from the English statement “Sarah is having a party this weekend,” the intonation pattern can be changed from falling at the end to rising at the end “Sarah is having a party this weekend?” (an echo question) or the word order can be changed to give “Is Sarah having a party this weekend?” (a yes/no question). American Sign Language (ASL), like some spoken languages, does not use the option of changing the word order but instead adds nonmanual markers. In this
example, the nonmanual marker is that of a “Yes/no question,” e.g., raising eyebrows. Such a marker is used to denote questions that can be readily answered with a simple “yes” or “no.” This is in contrast, for instance, to “Wh-questions” which start with a “wh”-word (or historical variant “h”) such as “which,” “when,” “how,” etc; in ASL Wh-questions are made with both the addition of nonmanual markers and optional word order changes. But each of these markers, for a “Yes/no question” or a “Wh-question,” may consist of multiple articulations, the most prominent being the position of the eyebrows, but with secondary articulations that may turn out to have their own meanings which combine with the primary meaning, or that may have emphaserizer effects on the primary meaning, or that may be signer-specific, or even accidental and irrelevant [157]. When these functions are combined with the possible articulations and efforts to generalize to signer-independent patterns, the problem quickly becomes intractable.

To identify nonmanual markers, linguists typically use computational tools to carefully annotate head movements and facial expression events in a set of sentences (Chapter 2). The analysis is then performed visually by identifying co-occurring nonmanuals events and grammatical markers, typically in small datasets. Clearly the complexity of this visual process becomes unmanageable when both the number of nonmanual events and samples increases. For this reason, it is imperative to design computational tools that can deal with such limitations.

In order to uncover the grammar of the face, we first test the hypothesis that an articulated model of the face would encode grammatically consistent and differential features. To that end, we manually annotated thousands of sentences jointly with a temporal logic approach to find discriminant nonmanual features from a set of videos. This approach involves two steps: first the procedure is validated by comparing the results with known discriminative features, that is, those already identified by sign language linguists, and then additional discriminative features and temporal structures are provided to linguists for further investigation and interpretation. This means that some features are known at the
outset, but most are uncovered by the computational algorithms defined in the present document. Then, a discriminant algorithm automatically identifies facial features that correlate with a grammatical marker but do not co-occur elsewhere.

To answer the question of the origin of some facial nonmanuals, we studied whether facial expressions of negative moral judgment (i.e., anger, disgust and contempt) have evolved into a facial expression of negation regularly used as a grammatical marker in human language (Chapter 3). Note that, we based aforementioned hypothesis on the fact that some facial expressions of emotion have evolved from the muscle development used in sensory regulation and later used to express moral judgment. In particular, we showed that this nonverbal signal is used as an alternative nonmanual of negation in ASL, meaning that this facial expressions has been grammaticalized.

Moreover, to alleviate manual marking process, we developed an automatic fiducial detection of the main components of the face i.e., eyebrows, eyes, mouth, chin and jawline, using salient and non-salient information of the face and a probabilist graphical model. To optimize the detection, we use Gibbs sampling to generate a set of possible detections, therefore automatically choosing the most likely one using the probabilistic graph (Chapter 4). We anticipate that an accurate detection of the components of the help would facilitate the detection of facial events.

Furthermore, we use modern computer vision and pattern recognition techniques to go beyond discrete representation of facial events in ASL sentences. This is done by designing new algorithms to deal with functions describing the movement of head and facial expressions (Chapter 4). In particular, we developed new pattern recognition techniques based on Support Vector Machines (SVM) that classify the continuous functions describing the variation of features over time, and most of the methods in this document can be applied to a variety of problems.
1.1 Literature Review

Recall that our main goal and contributions are not directly linked with the recognition of well defined grammatical sign language structures, instead, our contribution is designing state-of-the-art tools that allow us to find new grammatical markers. Unfortunately the computer vision community has been focused on designing automatic recognition systems, ignoring the need of a well defined grammar. Below, we summarize the state-of-the-art research for the study of sign language computer vision systems.

As mentioned before, most sign languages are composed of a manual part and nonmanual counterpart. As noted in [40], it is well known how the handshape, hand movement and palm orientation form the main building blocks of the manual component of the sign [140, 141, 14, 23, 134]. To date, the vast majority of automatic sign language recognition systems attempt to recognize recorded sentences using just the manual information. Even though, it is an active research topic, many state-of-the-art systems simply ignore the nonmanual components. A comprehensive review of recognition of manual sign language recognition can be found in [96, 120].

One of the most important problems with automatically recognizing sign language is the lack of full grammatical definition. Many nonmanuals give prosodic information, morphology and grammatical markers of different types of sentences [167, 12]. In the last 15 years the computer vision community has shifted its attention to recognize mainly known nonmanuals. Roughly speaking, automatic study of nonmanuals literature can be divided in three categories: a) interaction with manual counterparts, b) recognition of known nonmanuals and c) identification of grammatical nonmanual features.

With respect to interaction with manual counterparts, in [121, 136] a probabilistic approach is used to model the relationship between low-level features of the body parts. The resulting feature space is combined with the optical flow of the movement of the head to improve the recognition of sentences in continuous sign language. The authors stated that
using a simple representation of the movement of the head, can increase the recognition up to 4% in their experiments, especially in negative sentences, which are highly correlated with headshakes. In [153, 154], the authors combined hand tracking and Active Appearance Models (AAM) to define a vector space composed of head pose and second order facial configural information. Their method seems to slightly increase the recognition in German Sign Language (GSL). In [8], the authors model the time varying interaction using one Hidden Markov Model (HMM) for manuals and nonmanuals and an additional HMM for nonmanuals in order to identify ASL manual signs. Note that the aforementioned works do not identify nonmanual markers neither find interesting facial grammatical features; the nonmanual information is just used as additional features for recognition of manual signs.

Lately, automatically recognizing known nonmanuals is becoming an important topic in computer vision and sign language linguistics since some sentences can be identified purely based on their nonmanuals. For known nonmanuals, we refer to the few facial or head movement events that are highly correlated with a grammatical structure and that are generally accepted in the linguistic community of a particular sign language, e.g., in ASL lowering the eyebrows is related with Wh-questions. In [107] a 2D and 3D AAM-like algorithm is utilized to track the facial components and position of the head. After this step, the information is combined using Support Vector Machines and HMM to perform classification of five types of sentences using just nonmanual information. Similarly, in [95] a face shape detector is used to track changes in the face and head pose. A classifier is the used to detect low level features such as headshakes, headnods and brow raising in order to classify different types of construction using a linear chain Conditional Random Field (CRF). In [151, 152] a 3D deformable model is proposed in order to track four states of head rotation i.e., no rotation, start, apex and end of rotation. In [116] a facial tracker is used to find the state of head rotations and facial expressions to identify six known nonmanuals in ASL. For this purpose the authors defined low level primitives for features such as head shakes, brow raising and eye squint.
In summary, the before mentioned works track the facial features and the head rotations to identify previously discovered nonmanuals. The aforementioned methods find features that are directly correlated with a particular type of construction e.g., lowering of eyebrows with ASL Wh-questions, raising eyebrows in ASL Yes/no questions, among others.

Recently there have been a few works linking facial events and unknown nonmanuals. Note that discovering all nonmanuals will allow us to fully uncover sign language grammar, and it will allow us to better teach sign language as a second language. To the best of our knowledge, our work in [16] was the first paper using pattern recognition techniques and temporal logic features to find co-occurrence of facial and head events in different constructions (Chapter 2). In addition to our work, in [7] unsupervised learning of face and head rotation features are used to correlate linguistic events such as the beginning of a sentence and phrase structure in Greek Sign Language (GSL). Also, in [7] it was hypothesized that the mouth has an important role in sign languages, a hypothesis is partially corroborated in our work [16].
Chapter 2

Discriminant Features and Temporal Structure of Nonmanuals in American Sign Language

2.1 Summary

To fully define the grammar of American Sign Language (ASL), a linguistic model of its nonmanuals needs to be constructed. While significant progress has been made to understand the features defining ASL manuals, after years of research, much still needs to be done to uncover the discriminant nonmanual components. The major barrier to achieving this goal is the difficulty in correlating facial features and linguistic features, especially since these correlations may be temporally defined. For example, a facial feature (e.g., head moves down) occurring at the end of the movement of another facial feature (e.g., brows moves up), may specify a Hypothetical conditional, but only if this time relationship is maintained. In other instances, the single occurrence of a movement (e.g., brows move up) can be indicative of the same grammatical construction. In the present chapter, we introduce a linguistic–computational approach to efficiently carry out this analysis. First, a linguistic model of the face is used to manually annotate a very large set of 2,347 videos of ASL nonmanuals (including tens of thousands of frames). Second, a computational approach is used to determine which features of the linguistic model are more informative of the grammatical rules under study. We used the proposed approach to study five types of sentences – Hypothetical conditionals, Yes/no questions, Wh-questions, Wh-questions postposed, and Assertions – plus their polarities – positive and negative. Our results verify
several components of the standard model of ASL nonmanuals and, most importantly, identify several previously unreported features and their temporal relationship. Notably, our results uncovered a complex interaction between head position and mouth shape. These findings define some temporal structures of ASL nonmanuals not previously detected by other approaches.

2.2 Introduction

To identify nonmanual markers, sign language researchers will typically manually annotate head movements and facial expression changes observed in a large number of video sequences. Tools such as ELAN [26] have been specifically designed for this purpose, Fig. 2.1. ELAN allows visual observation of the starting and ending frame of the video sequence for each of these manual annotations. Furthermore, ELAN is a powerful tool that allows extracting data depending on the tiers, signed sentences, type of clauses or references over an interval of time among others. However, the aforementioned tool is not designed to perform statistical analysis and pattern recognition algorithms over the previously manually marked data. For this reason, analysis about the annotations is typically performed through a careful visual analysis to identify co-occurring nonmanuals and grammatical markers in large numbers of video sequences.

To date, research in ASL has identified that Hypothetical conditionals, Yes/no questions and Wh-questions are marked primarily by nonmanuals and secondarily by optional signs (e.g., for conditionals, a sign with the meaning ‘if’ may be used but is not required) [11]. It has also been hypothesized that Wh-questions which have word order changed with the Wh-word moved to the end (“postposed”) could involve other distinct nonmanuals than those than those used in ordinary Wh-question [32, 33, 3]. Moreover, polarity (i.e., positive versus negative) seems to be marked with nonmanuals; there is no regular sign for indicating positive polarity as this is the default interpretation in all languages, and negative signs for
Figure 2.1: ELAN is a computer software that allows users to view synchronized videos simultaneously and frame by frame (top of figure), facilitating manual annotations (bottom half). Our manual annotations specify where the sentence starts and ends, where each word (or concept) starts and ends, plus the shape and configural features used to uncover the linguistic model (see text).

Negative polarity are optional if the nonmanual for negation is present [148, 111, 123]. Due to the slowness of the standard approach used by linguists, it is difficult to verify to what extent these results hold over a larger number of video sequences or signers. Thus, it is unclear whether these are the only (required) nonmanuals used in these sentence types.

The present chapter describes a linguistic-computational approach to automatically finding discriminant nonmanual features from a set of annotated videos. This approach involves
two steps: first the procedure is validated by comparing the results with known discriminative features, that is, those already identified by sign language linguists, and then additional discriminative features and temporal structures are provided to linguists for further investigation and interpretation. This means that some features are known at the outset, but most are uncovered by the computational algorithms defined in the present chapter. Taken together, these discriminant features and temporal structures comprise an expanded linguistic model of the nonmanuals under study. To achieve this goal, videos are annotated using a linguistic/articulated model of the face. Then, a computer algorithm automatically identifies facial articulations that correlate with a grammatical marker but do not co-occur elsewhere. The algorithm finds single nonmanual markers, such as a single facial component (e.g., brows up), and first-order co-occurrences (i.e., temporal structure), as, for example, one facial or head articulation occurring before another (e.g., head turns right before brows move up). Note that the term “discriminant” goes beyond a characterization of the nonmanual. While characterization defines the production of a nonmanual, discriminant features are those produced during one grammatical construction (e.g., wh-question) but absent elsewhere. This proposed approach will thus be used to test the hypothesis that nonmanual markers discriminate among the following nine classes of sentences: Hypothetical conditionals, Yes/no questions, Wh-questions, Wh-questions postposed, Assertions and their polarities (positive and negative).

This proposed approach not only validates some known nonmanuals but, most importantly, identifies a large variety of previously unsuspected nonmanual markers for each of the nine sentence types of ASL considered in the present chapter. For example, as expected, our results show a systematic relationship between eyebrow position and grammatical constructions. As predicted by previous literature, ‘brows move up’ is prominent in Hypothetical conditionals (89.1%) and Yes/no questions (92.3%). Similarly, ‘brows move down’ occurs systematically in Wh-questions (89.5%) and Wh-questions with the Wh-sign postposed (99.2%). However, our results reveal a complex interaction between head position
and mouth shape that has not been previously reported in the literature. This finding is extremely relevant because it shows how co-articulations of facial components are employed as grammatical constructions and hence emphasizes the importance of complex interaction of nonmanual markers in sign language.

The results summarized in the preceding paragraph would have been difficult to attain using a visual analysis of manual annotations. In contrast, the proposed computational approach can search for all possible first-order feature relationships and calculate which consistently co-occur in a given grammatical construct but rarely happen elsewhere. The approach and algorithms described in this chapter have been incorporated into ELAN and can hence be readily used by other researchers to replicate and expand on the results reported herein.

### 2.3 Methodology

We investigate the role of nonmanuals in five (5) types of sentences: Hypothetical conditionals, Yes/no questions, Wh-questions, Wh-questions postposed and Assertions; in addition we consider their polarities: positive and negative. This yields a total of 9 classes because Yes/no-questions are neutral, meaning they cannot be associated with a specific polarity (although this does occur in some other languages, *e.g.*, spoken English and Turkish Sign Language both allow negative Yes/no questions [131, 61]).
2.3.1 Database

We recorded fifteen (15) Deaf native users of ASL signing more than 129 distinct sentences each [164]. Each of these sentences corresponds to the 9 classes (Section 2.6) defined above (i.e., Hypothetical conditionals, Yes/no questions, Wh-questions, Wh-questions postposed, Assertions, and their polarities), for a total of 2,347 video sequences, although for variety of targets, not every signer produced exactly the same set of stimuli to incorporate variability in the data. This data variability is key to find generalizations of the model. For instance, we wish to see if the same discriminant temporal correlates are found in similar linguistic structures even when the productions differ; see Section 2.6 for lists of stimuli. It should be noted that signers were asked to replicate a series of sentences after watching video recordings of them. In this case, signers do not replicate the sentence (or group of sentences) exactly as in the video, but its meaning. Subject variability is expect and is indeed present in the collected dataset as was made clear after a careful analysis of each video sequence. Note that our goal is to use data with sufficient variability to allow us to recover the computational model of nonmanuals. This model can be put into test in subsequent field studies.

The signers were recorded using two high quality Sony DCR-VX2100 cameras. These cameras are equipped with 3 1/3” CCDs for fast capture of color images in our studio conditions. All human subjects signed a consent form, granting permission for the use of their video sequences in research and the replication of these in scientific articles. The research and consent forms were approved by the IRB boards at The Ohio State University and Purdue University.

The first camera recorded the upper-body (including the head) of the signer. The second camera captured a close-up of the face. This second camera provides high-quality video of the nonmanuals, Fig. 2.2. Watching both videos together, the sign language researchers manually labeled each video sequence as belonging to one of the five types of sentences listed in Section 2.6 and to one of their polarities. The sentences we consider are in Tables
S1-S4 and the sentences signed by each one of the 15 participants in our database are in Table S5. These sentences correspond to 506 Hypothetical conditionals, 350 Wh-questions, 124 Wh-questions postposed, 313 Yes/no-questions, and 1,054 Assertions.

For consistency check, the annotations of each recorded sentence were visually validated by a native Deaf ASL signer and an experienced sign language researcher who were members of the American Sign Language Linguistics Laboratory at Purdue University. In particular, we made sure all video clips in the database correctly expressed its target sentence and that it was clearly visible and understood. Video clips not passing this test were eliminated from the database.

The video clips and manual annotations described in this section will be made publicly available to those wishing to extend on the results reported herein.

2.3.2 Manual Annotations

Research in face perception has demonstrated that facial expressions are coded and recognized by the cognitive system using configural [113] and shape [114] features. Configural refers to second-order changes. First-order changes code for the ordering of features (e.g., nose on top of the mouth), while second-order specify between-feature distances. Shape features means that facial features are in a specified position (e.g., the curvature of the mouth). These descriptions are correlated with facial movement that may also be defined using other coding systems [38].

Similarly, sign language research has shown that such options as brow position, closed/open mouth and flat/round lips, teeth showing, and head turns are potential building blocks of nonmanual markers [167, 22]. We thus used fifteen (15) configural and shape feature positions corresponding to each of these nonmanual building blocks to annotate facial expressions in the video sequences of our database. These fifteen labels are summarized in Table 2.1 and Fig. 2.3.
All video clips are displayed with the ELAN [26] software. A benefit of the ELAN software is that video sequences can be displayed frame by frame in synch with a time cursor so that the desired location for an event can be identified. A sign language expert can then manually annotate the configural and shape positions described above. This means that each annotation specifies where a configural or shape position starts and ends. An example of such a manual annotation is shown in Fig. 2.1. The manual annotations were reviewed by the two Purdue co-authors and, if necessary, changes were made until there was agreement in the coding.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move</td>
<td>{Up, Down}</td>
</tr>
<tr>
<td>Blinks</td>
<td>{Blink}</td>
</tr>
<tr>
<td>Mouth</td>
<td>{Open, Closed}</td>
</tr>
<tr>
<td>Mouth shape</td>
<td>{Round, Flat, Other}</td>
</tr>
<tr>
<td>Teeth</td>
<td>{Closed, Open, Touch lip}</td>
</tr>
<tr>
<td>Head turns</td>
<td>{Left, Right}</td>
</tr>
<tr>
<td>Head tilts</td>
<td>{Left, Right}</td>
</tr>
<tr>
<td>Head moves</td>
<td>{Up, Down}</td>
</tr>
</tbody>
</table>

The qualitative manual annotations described above must then be quantified in order to determine the most discriminative facial features. A possible solution is to treat a feature as a time varying function, where each category has some numerical value [126, 81]. The problem with this approach is that the sentences need to be aligned, that is, they must be shrunk or expanded to a canonical length. This would diminish or overemphasize some feature categories, especially those that expand a shorter time interval. Moreover, this approach would not model sequences of events, e.g., headshakes, left to right turns, etc. We resolve these problems using Allen’s Temporal Logic (ATL).
Figure 2.3: The configural and shape positions used to define each of the nonmanuals in sentences of ASL. In 2.3(a) we show a neutral face. A neutral face is defined as one without expression where all facial muscles are relaxed (except for the eyelids which are open). 2.3(b) We consider two configural positions for the eyebrows (up and down). 2.3(c) Blinks are marked by closing the eyelids. 2.3(d) The mouth can be open or closed. 2.3(e) We also annotate mouth shape where appropriate (flat, round and other). 2.3(f) When there is teeth showing, we consider three distinct positions – closed (top and bottom teeth touching), open (not touching), touching lips (where the top teeth are over the lower lip or the bottom teeth touch the upper lip). 2.3(g)-2.3(i) We also consider the three possible rotations of the head – turns, tilts and forward/backward movements.
2.3.3 Temporal Logic Description

ATL is a framework that allows us to analyze relative temporal information, such as event $A$ happens before event $B$ [5]. Here, any two time events are related by a set of symmetric, mutually exclusive binary relations, called propositions. In our modeling, we employ the following set of propositions: before, meets, overlaps, equals, starts, during and finishes. To show the use of the above defined propositions, consider the examples in Fig. 2.4. In this figure, we have two events, $A$ and $B$. $A$ is said to be before $B$, when $A$ happens disjointly before $B$, Fig. 2.4(a). For example, $A$ could be head turns right and $B$ head turns left. Here, we would write head turns right before head turns left. This could be the case when a subject is signing a negative statement with negation marked with a headshake.

In the case that $A$ happens immediately before $B$, then $A$ is said to meet $B$, Fig. 2.4(b). Note that the difference between before and meets is that before requires a non-empty time interval between both events. For example, when nodding, the head moves up and down without a visual pause, which could be written as, $A$ meets $B$. Obviously, in practice, two events involving different articulations would only perfectly follow one another by chance. To accommodate for small natural variabilities (e.g., those due to data acquisition or small variations of the natural human movement between different subjects), we define meets as $B$ occurring after a very brief interval $\delta t$ after $A$. The value of $\delta t$ will be estimated using cross-validation in learning. In cross-validation, we divide the training data into two or more sets; use all but one of those sets for training while using the left out set for testing values of $\delta t \in [0, \epsilon]$, with $\epsilon$ small. This is repeated multiple times to determine the value of the parameter yielding better generalizations. This is a common practice in pattern recognition where a learning algorithm uses a training set to come up with a representation that accurately represents some observations or discriminates between observations belonging to different categories (classes). A testing set is then used to determine whether the learned representation is capable of discriminating previously unseen examples into the correct class.

$A$ is said to overlap $B$ when $A$ starts before $B$ and $A$ finishes during $B$, Fig. 2.4(c). In
contrast, equals means that both events, $A$ and $B$, share the same time interval, Fig. 2.4(d).

This proposition is useful to denote single featural events, e.g., to indicate that the brows move up once, as in Yes/no questions [11]. Although this may seem redundant at first, this notation allows us to consider single actions without changing notation or the algorithm.

When both events start at the same time but $A$ finishes before $B$, then $A$ is said to start with $B$, Fig. 2.4(e). Similarly, when events $A$ and $B$ finish at the same time but $A$ starts after $B$, then $A$ is said to finish at $B$, Fig. 2.4(f). Finally, during means that $A$’s time interval happens within $B$’s time interval, Fig. 2.4(g).

Fig. 2.5 shows an equivalent time diagram for the manual annotation previously illustrated in Fig. 2.1 for the sentence “#BRAD-IXi COOK FISH ON GRILL IXi,” (i.e., “Brad is cooking/cooks fish on the grill”). The resulting coding using ATL relations is shown in Table 2.2.

In summary, the Allen’s Temporal Logic defined above is composed of a set of binary propositions. Formally, we denote this set as $\mathcal{P} = \{\text{before, meets, overlaps, starts, during, finishes}\}$. The set $\mathcal{P}$ operates over the time interval defined by the set of events $\mathcal{I}$. Therefore, an ATL can be formally denoted as $\text{ATL}(\mathcal{P}, \mathcal{I})$. In this notation, any two events $i, j \in \mathcal{I}$ are related using one of the propositions in $\mathcal{P}$, e.g., before$(i, j)$ specifies that event $i$ happened before event $j$. 
Table 2.2: Temporal relations for the events in Fig. 2.5. It is important to note in the table below that the temporal information is encoded in the description of the ATL using the propositions in $P$. The temporal information is hence intrinsically coded in this table.

<table>
<thead>
<tr>
<th>Event</th>
<th>Before</th>
<th>Meets</th>
<th>Overlaps</th>
<th>Equal</th>
<th>During</th>
<th>Starts</th>
<th>Finishes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head turns left</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head tilts left</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head turns right</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head moves down</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Blink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Head moves up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

19
Figure 2.5: A diagram describing facial feature movements/positions and the gloss for the sentence “#BRAD-IXi COOK FISH ON GRILL IXi.” For clarity, here, we have listed the events and their time intervals in order of occurrence. The top row specifies the first event, with subsequent rows listing later occurring events. The bottom row summarizes the time interval of each signed concept. This visualization facilitates the coding of the events using the propositions in $P$. For example, for the figure above, it is easy to see that head moves up occurs during event brows move up, which can be compactly expressed as during(head moves up, brows move up).

The 17 feature categories (Table 2.1) form a set of 2,023 possible ATL first-order relations. We eliminated relations that cannot co-occur due to their mutually exclusive nature, (e.g., brows move up equals brows move down) giving a total $d = 1,789$ feasible relations.

It is also important to encode the number of consecutive occurrences for a given ATL relation. This might be important for some discriminant features, e.g., while a single head-shake may not carry any grammatical meaning, multiple headshakes can be a marker of negation or Wh-questions [157]. To correctly represent this information, we encode the relative frequency of each occurrence in a histogram, which displays the number of times
Figure 2.6: Visual representation of the ATL feature vector \( \mathbf{x} \). The dark blue color indicates a low number of occurrences for an event, while a dark orange color indicates a high number of repetitions. This figure is the histogram corresponding to the example in Fig. 2.5. The feature vector entries \( (x_k) \) are read from left to right (\( k = 1, \ldots, 49 \)).

That a given event happens.

Formally, we represent a sentence as

\[
\mathbf{x} = (x_1, \ldots, x_d)^T,
\]

where \( x_k \) is the number of times that the first-order relation \( 1 \leq k \leq d \) repeats in a sentence. For instance, if a sentence includes four eye blinks, the feature vector \( \mathbf{x} \) will have a value of 4 in the position \( x_{\text{blinks}} \); where we have used \( k = \text{blinks} \) to indicate that this is the feature used to code for blinks. Fig. 2.6 shows the histogram for the example previously shown in Table 2.2 and Fig. 2.5.

### 2.3.4 Discriminant Analysis

The histogram representation of the ALT described thus far provides a convenient numerical representation of the nonmanual events we wish to study. To determine the time relations that best discriminate a grammatical structure from the rest (e.g., Yes/no-questions versus the others), we need to use a feature extraction algorithm that uncovers the features or combinations of them that best discriminate between sentence types. In pattern recognition, such approaches are called discriminant analysis [103]. When the number of samples (relative to the number of features) is small, as is the case in the present study, Regularized Linear Discriminant Analysis (RLDA) is a possible algorithm to use [55]. RLDA adds a regularizing factor to the metrics being computed, preventing singularities even when the
number of samples is small or when the underlying metric cannot be fully estimated \[178\]. Also, RLDA has a single parameter to estimate, making it very efficient and easy to work with \[55\].

Formally, RLDA finds the projection vector $w$ that best separates (in the least-square sense) two classes by maximizing the ratio between the class means to the average variance of these classes. Consider the case where $C_1$ and $C_2$ represent class 1 and 2, respectively. And, let the sample sets be $\{x_{i1}, \ldots, x_{in_i}\}$, where $i$ specifies the class and $n_i$ the number of samples belonging to it. The discriminant hyperplane separating the samples of these two classes is defined by its normal vector, $w$. This vector is given by,

$$w^* = \arg \max_w \frac{\|w^T(\mu_1 - \mu_2)\|^2_2}{w^T(S_W + \lambda I)w}, \quad s.t. \quad \|w\|_2 = 1,$$

(2.1)

where $\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$ are the sample class means, $S_W = \frac{1}{2} \sum_{i=1}^2 \frac{1}{n_i} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^T$ is the sample within-class scatter matrix, $\lambda$ is the regularizing parameter that is found using cross-validation, $I$ is the identity matrix and $\|\cdot\|_2$ specifies the 2-norm (euclidean) measure. Recall that the regularizing parameter is used to ensure the above equation has a robust solution when the number of samples is small (i.e., even if the within-class scatter matrix is singular).

Solving for (2.1) yields, $w = (S_W + \lambda I)^{-1}(\mu_1 - \mu_2)$.

An ATL relation is hence defined as discriminative if its corresponding absolute magnitude in $w$ is larger than the others i.e., $|w_i| > |w_j|, \forall i \neq j$. To rank their relative importance, each element of the vector $w$ is normalized with respect to its largest attained value, i.e., $\tilde{w} = \frac{|w|}{\|w\|_\infty}$ with elements $\tilde{w}_i \in [0,1]$, with 0 meaning the worst possible feature and 1 meaning the most important one, and where $|w| = (|w_1|, \ldots, |w_d|)$, and $\|w\|_\infty = \max_{1 \leq i \leq d} |w_i|$.

Our hypothesis is that nonmanual markers can be used to discriminate among the nine classes of sentences described above. More specifically, we hypothesize that first-order temporal relations of facial movements are sufficient to code for such grammatical structure.
To test this hypothesis, we use all the video sequences in our database except one to find the discriminant facial features (as described in the Methods section) and test whether the resulting model correctly classifies the left out sentence. This approached is known as Leave-One-Sentence-Out (LOSO) test.

Classification of the left out (test) sample $\mathbf{x}_{test}$ is done using the nearest-mean classifier. The nearest-mean classifier assigns to $\mathbf{x}_{test}$ the class label $i$ of the nearest class mean $\mu_i$, i.e., $T(\mathbf{x}_{test}) = \text{argmin}_i \| \mathbf{w}^T (\mu_i - \mathbf{x}_{test}) \|_2$. If we have $n$ sample signed sentences, there are $n$ possible sentences we can leave out in the LOSO approach. In LOSO, we try all these $n$ possibilities and then compute the mean classification accuracy. We also estimate the expected $\tilde{\mathbf{w}}$ by averaging the $\tilde{\mathbf{w}}$ vectors generated from all LOSO iterations. Note that we only compute the classification accuracy for the features that provide the largest $\tilde{\mathbf{w}}$, since this value is correlated with discriminability.

In addition to the above, we included the commonly used sensitivity index $d'$ to measure the distance between signal and noise for the most discriminative features. Here, $d'$ measures the performance of a single feature in isolation and, hence, does not provide information on co-occurring features or their temporal structures.

### 2.4 Results

#### 2.4.1 Experiment 1: Constructions Discriminant Features

First, we wish to determine the nonmanuals that best discriminate each structure, i.e., the discriminant features. To achieve this, we run a one-versus-all experiment. This means that, for each class (e.g., Wh-questions), we use the linguistic-computational approach described in the Methods section to find the discriminant features that are common to that class but are not descriptive of the other classes.

The resulting discriminant features need to distinguish between the grammatical structures under study. These features are those providing the highest classification accuracies in
the LOSO test described above. They are in Tables 2.3-2.7. The two columns in these tables labeled “% Activation” specify the characterization of the nonmanuals, i.e., the number of times the nonmanual is employed to marked a grammatical construction.

In Tables 2.3-2.7 we also specify the classification accuracy of each of the discriminant features found with the proposed approach. To do this we use the following approach. Each discriminant feature $f_k$ defines a one-dimensional feature space $F_k$ with its corresponding basis vector $f_k$. We project all vectors $x_i$ onto $F_k$, i.e., $x_i^T f_k$. We then use RLDA to learn the hyperplane $h_k$ that best separates the samples of our two classes. Note that Linear discriminant analysis and RLDA provide the Bayes optimal solution when we have only two classes with equal variances [103]. Once this hyperplane has been determined, we compute the percentage of samples belonging to class 1 (i.e., $x_i \in C_1$) that are on one side of $h_k$ and the percentage of samples of class 2 (i.e., $x_i \in C_2$) that are on the other side. These two numbers provide the percentage of classification accuracies listed in the last two columns in Tables 2.3-2.7.

The numbers in these last two columns (labeled “% Classification”) specify how many of our sentences can be correctly classified using each single feature $f_k$. This refers to how discriminant the feature is. Some discriminant features will of course be more common and, hence, will successfully discriminate more samples of $C_j$ than others, with $j = \{1, 2\}$. For example, “Head moves down finishes brows move up” in Table 2.3 is not a common nonmanual marker for Hypothetical conditionals (only used in 19% of Hypothetical conditionals), but it is almost never used elsewhere (2.3% of other sentences). This makes it a very efficient, robust stand-alone nonmanual to indicate a sentence is not a Hypothetical conditional (with classification accuracy at 97.6%). In comparison, “Brows move up” is a better nonmanual marker of conditionals (since 89.1% of our Hypothetical conditionals are successfully classified with it), but is also employed elsewhere (46.3% of other sentences are also classified as Hypothetical conditionals if one were to use only this feature). Thus, this second nonmanual is not as robust as the previous one. As expected, the result of averaging...
Table 2.3: Discriminant features for Hypothetical conditionals.

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>$\tilde{w}_i$</th>
<th>% Activation in Conditionals</th>
<th>% Activation in Others</th>
<th>d’</th>
<th>% Classification in Conditionals</th>
<th>% Classification in Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move up</td>
<td>1</td>
<td>89.1</td>
<td>54</td>
<td>1.13</td>
<td>89.1</td>
<td>46</td>
</tr>
<tr>
<td>Head moves down finishes brows move up</td>
<td>0.78</td>
<td>19</td>
<td>2.3</td>
<td>1.12</td>
<td>19</td>
<td>97.6</td>
</tr>
<tr>
<td>Head turns left during brows move up</td>
<td>0.67</td>
<td>50.4</td>
<td>13.6</td>
<td>1.11</td>
<td>50.4</td>
<td>85.2</td>
</tr>
<tr>
<td>Brows move up equals head moves down</td>
<td>0.65</td>
<td>18</td>
<td>2</td>
<td>1.14</td>
<td>18</td>
<td>97.9</td>
</tr>
<tr>
<td>Teeth touch lip during brows moves up</td>
<td>0.64</td>
<td>41.3</td>
<td>6.4</td>
<td>1.31</td>
<td>41.3</td>
<td>93.7</td>
</tr>
<tr>
<td>Mouth shape other equals mouth open</td>
<td>0.58</td>
<td>69</td>
<td>53</td>
<td>0.42</td>
<td>34.4</td>
<td>77.2</td>
</tr>
<tr>
<td>Teeth open</td>
<td>0.56</td>
<td>92.1</td>
<td>84.4</td>
<td>0.40</td>
<td>61.2</td>
<td>68</td>
</tr>
<tr>
<td>Mouth open equals</td>
<td>0.55</td>
<td>11.5</td>
<td>2.5</td>
<td>0.76</td>
<td>11.5</td>
<td>97.6</td>
</tr>
<tr>
<td>brows move up</td>
<td>0.54</td>
<td>37.2</td>
<td>13.9</td>
<td>0.76</td>
<td>37.2</td>
<td>87.7</td>
</tr>
<tr>
<td>Teeth touch lip during mouth open</td>
<td>0.53</td>
<td>87</td>
<td>77.1</td>
<td>0.38</td>
<td>73.9</td>
<td>30.9</td>
</tr>
</tbody>
</table>

the last two columns in Tables 2.3-2.7 are highly correlated with d’. This is because both methods of analysis assume the data is Normally distributed. This correlation however is stronger for the single feature case, since d’ cannot account for temporal structure.

Additionally, we tested for the statistical significance of our results. This was done by comparing our results with those given by a randomization of the class labels. That is, we compare the results obtained with the proposed approach to the results one observes when the class labels for each of the samples $x_{ij}$ are assigned to a random class (rather than their true class label $i$). The randomization was repeated 24 times, yielding a total of 24 classification results. These results specify the probability of obtaining the classification accuracies by chance. A $t$-test of these revealed that our method performed significantly better than chance with the following $p$ values: $p < 10^{-27}$ for Hypothetical conditionals, $p < 10^{-32}$ for Wh-question, $p < 10^{-21}$ for Wh-questions postposed, $p < 10^{-28}$ for Yes/no questions and $p < 10^{-24}$ for Assertions.

Let us now describe the results of this study in detail for each of the 5 classes under consideration.
Hypothetical Conditionals

With respect to the Hypothetical conditionals (Table 2.3), the high percentage of “brows move up” is expected from the literature [11, 12, 91, 167, 163], as the conditional clause is routinely marked by raised brows. However, within the conditional clause, individual signs may require another facial posture that interferes with raised brows [158], and therefore not every sign in a Hypothetical conditional will have raised brows marked on it, thereby accounting for the less than 100% occurrence. For example, a facial expression that could interfere with the marking of conditional might be that of surprise, which involves brows up, head back, and eyes wide open. Furthermore, within the structures that are not Hypothetical conditionals (fourth column Table 2.3), there are Yes/no questions and topics in Assertions, which also are routinely marked by raised brows. Thus, 54% of the non-Hypothetical conditionals also show “brows move up.”

Most notably, Table 2.3 provides novel (and some unexpected) results concerning the behavior of the head, and the mouth and teeth. For instance “head moves down finishes brows move up” in 19% of the Hypothetical conditionals suggests a head thrust at the end of the conditional clause [92] and/or a prosodic reset [31, 32, 33] prior to the onset of the clause following the Hypothetical conditional clause.

Another frequent head behavior is “head turns left during brows move up,” which may reflect the establishment of a space to the left of the signer at head level to mark clauses containing content that is uncertain, hypothetical, or otherwise irrealis. The use of space for linguistic pragmatic functions has been recently reported for Catalan Sign Language (LSC) [13] and for Austrian Sign Language (OGS) [86]. Most relevant to the “head turns left during brows move up” in Hypothetical conditionals is Lackner’s observation of the signers’ reference to a “mental” space or “space of thoughts,” which may be coded by pointing, gazing up, or moving the chin up.

An additional head behavior, “head turns right,” raises another possible interpretation for “head turns left” in conditionals. As will be discussed in the Polarity section below,
“head turns right during brows move up” occurs very frequently in clauses containing negation (negative polarity), as part of the negative headshake (right-left-right sequences [10]). Thus, the frequent occurrence (50.4%) of “head turns left during brows move up” in Hypothetical conditionals is highly associated with negation.

Both Hypothetical conditionals and non-Hypotheticals have a high occurrence of “teeth open” in Table 2.3. For the Hypothetical conditionals, this is likely related to the frequent articulation of the word “if” when the sign IF is produced. This suggestion is strengthened by the more frequent occurrence of “teeth touch lip during brows move up” and “teeth touch lip during mouth open” in Hypothetical conditionals than in non-Hypotheticals – i.e., the (upper) teeth touch the (bottom) lip at the end of the articulation of “if.” In contrast, the high frequency of “teeth open” in non-Hypotheticals is not accompanied by high occurrence of “teeth touch lip during brows move up” and “teeth touch lip during mouth open.” Instead, “teeth open” is the result of the inclusion of lexical items such as FISH in the stimuli. As reported in [110], nouns in ASL and other sign languages are much more likely to be accompanied by mouthing of the surrounding spoken language word than other word categories (e.g., pronouns, verbs). Thus, it is not unusual that a noun sign like FISH would be accompanied by the articulation of “fish” or at least the first part of it that involves articulation of “f” or “fi.” Fig. 2.7 illustrates this in a sequence of mouth positions in one Wh-question produced by one of the ASL signers in our database.
Wh-Questions

Table 2.4: Discriminant features for Wh-questions.

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>$\bar{w}_i$</th>
<th>% Activation in Wh questions</th>
<th>% Activation in Others</th>
<th>$d'$</th>
<th>% Classification in Wh questions</th>
<th>% Classification in Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move up</td>
<td>1</td>
<td>10.6</td>
<td>70.5</td>
<td>1.79</td>
<td>89.4</td>
<td>73.6</td>
</tr>
<tr>
<td>Brows move down</td>
<td>0.99</td>
<td>89.4</td>
<td>23.2</td>
<td>1.98</td>
<td>89.4</td>
<td>62.4</td>
</tr>
<tr>
<td>Mouth shape round</td>
<td>0.7</td>
<td>43.1</td>
<td>0.5</td>
<td>2.44</td>
<td>43.1</td>
<td>99.3</td>
</tr>
<tr>
<td>starts brows move down</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouth shape flat</td>
<td>0.63</td>
<td>67.1</td>
<td>92.4</td>
<td>0.99</td>
<td>67.4</td>
<td>62.7</td>
</tr>
<tr>
<td>Mouth shape round</td>
<td>0.56</td>
<td>82.9</td>
<td>46.7</td>
<td>1.03</td>
<td>40.9</td>
<td>71.4</td>
</tr>
<tr>
<td>Head turns right starts</td>
<td>0.53</td>
<td>32</td>
<td>3.3</td>
<td>1.38</td>
<td>32</td>
<td>96.2</td>
</tr>
<tr>
<td>brows move down</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 2.4, we see that Wh-questions are separated from other constructions by both “brows move up” and “brows move down,” but in different ways. “Brows move down” is a well-known discriminant feature for Wh-questions in ASL [111, 12] and occurs in 89.4% of the Wh-questions in our sample. The occurrence of “brows move down” in 23.2% of the other constructions is likely related to the occurrence in those constructions of Wh-questions with the Wh-sign postposed (discussed separately in connection with Table 2.5). This is diminished when the downward movement of the brows is preceded by the head turning right.

In contrast, “brows move up” occurs in few Wh-questions (10.6%) but is very frequent in other constructions (70.5%), which includes the Hypothetical conditionals discussed above and Yes/no questions (discussed below), both of which are associated with raised brows. “Brow move up” may also be associated with some occurrences of Wh-questions with Wh-sign postposed. This allows for very high classification rates of Wh-questions and other constructions even when they are using this single feature.

The remaining discriminative cue is “mouth shape round starts brows move down,” which occurs frequently in Wh-questions (43.1%) but not in other constructions (0.5%). This cue is likely associated with the presence of mouthing of “who” at the beginning of some Wh-questions. This is also the case for “mouth shape round.”
From the results in Table 2.4, we can thus identify a primary cue “brows move down” and a secondary cue “mouth shape round starts brows move down” for Wh-questions.

**Wh-Questions postposed**

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>( \tilde{w}_i )</th>
<th>% Activation in Wh questions postposed</th>
<th>% Activation in Others</th>
<th>( \text{d}' )</th>
<th>% Classification in Wh questions postposed</th>
<th>% Classification in Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move down</td>
<td>1</td>
<td>99.2</td>
<td>29.4</td>
<td>2.95</td>
<td>99.2</td>
<td>64.7</td>
</tr>
<tr>
<td>Mouth shape other overlaps brows move down</td>
<td>0.68</td>
<td>43.5</td>
<td>4.9</td>
<td>1.50</td>
<td>43.5</td>
<td>94.6</td>
</tr>
<tr>
<td>Brows move down finishes mouth open</td>
<td>0.59</td>
<td>21.8</td>
<td>4</td>
<td>0.97</td>
<td>21.8</td>
<td>95.4</td>
</tr>
<tr>
<td>Blink overlaps brows move down</td>
<td>0.54</td>
<td>16.9</td>
<td>2.5</td>
<td>1.01</td>
<td>16.9</td>
<td>97.8</td>
</tr>
<tr>
<td>Mouth shape round during brows move down</td>
<td>0.5</td>
<td>61.3</td>
<td>9.5</td>
<td>1.6</td>
<td>61.3</td>
<td>87.2</td>
</tr>
<tr>
<td>Brows move up meets brows move down</td>
<td>0.49</td>
<td>37.1</td>
<td>5.1</td>
<td>1.30</td>
<td>37.1</td>
<td>94.8</td>
</tr>
</tbody>
</table>

A “Wh-question postposed” is one in which the Wh-word has been produced at the end of the question instead of at the beginning (described as “focus questions” in [31]). This placement of the Wh-word has the effect of allowing the main clause to be treated either as part of the question or as an Assertion followed by a question [112]. As a result, “brows move down” may cover the entire question or only the final Wh-word; either way, “brows move down” is a distinctive marker; Table 2.5. The occurrence of “brows move down” in other constructions is due to the inclusion of regular Wh-questions discussed above. When the signs preceding the postposed Wh-sign are treated as separate from the question at the end, we see very frequent (37.1%) “brows move up meets brows move down,” with the brows up on the non-question part and the brows down on the Wh-word. This “brows up meets brows down” pattern in ASL Wh-questions postposed is noted in [157] and discussed with respect to the presuppositional nature of the material preceding the postposed Wh-word in [3].

The mouth is also active in relation to “brows move down,” with “mouth shape round during brows move down” occurring in 61.3% of the Wh-questions postposed, as compared...
to only 9.5% in other constructions. Again, it is likely due to mouthing of “who,” which occurs frequently in Wh-questions postposed and also in regular Wh-questions which are included in the comparison constructions. “Mouth shape other overlaps with brows move down” frequently (43.5%) in Wh-questions postposed, and may be related to mouthing of other Wh-words, such as “which,” “why,” and “where.” Note the classification rate, for Wh-questions postposed and others is 94% when combining the features.

One articulation in Wh-questions postposed that did not show up in other constructions is the occurrence of blinks. “Blink overlaps brows move down” occurred in 16.9% of these as compared to only 2.5% in other constructions. Periodic blinks, the kind that are associated with eye-wetting, are well-known as a marker of the end of intonational phrases and syntactic constituents in ASL [160]. But if these blinks were just periodic blinks, they would occur after the brows move down ends. The fact that we see blinks overlapping with brows move down implies that they are deliberate blinks – slower and longer in duration. Deliberate blinks are associated with prominence on a sign [160]. If the blink ended at the same time as the brows move down, we would also know that the blink occurred on the last sign in the clause. The fact that blinks overlap with brows down means that the blink is located on a sign inside the clause. This supports the suggestion that they are deliberate blinks, which are used to emphasize a sign, because signs in final position in a clause are already emphasized/stressed [160] and therefore would not need a deliberate blink as a marker.

**Yes/no Questions**

Yes/no questions are distinguished primarily by “brows move up,” although this cue also occurs frequently in other constructions, which include Hypothetical conditionals (Table 2.6) and Assertions with marked topics. “Brows move up” and “brows move down” achieve very high classification accuracies for Yes/no questions – over 92%.

Note that, as expected, “brows move up before brows move down” does not occur in Yes/no questions, since the brow raise is expected to span the entire question [12]. In
Table 2.6: Discriminant features for Yes/no-questions.

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>$\tilde{w}_i$</th>
<th>% Activation in Yes/no questions</th>
<th>% Activation in Others</th>
<th>$d'$</th>
<th>% Classification in Yes/no questions</th>
<th>% Classification in Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move down</td>
<td>1</td>
<td>6.1</td>
<td>37.3</td>
<td>1.22</td>
<td>93.9</td>
<td>58.6</td>
</tr>
<tr>
<td>Brows move up</td>
<td>0.81</td>
<td>92.3</td>
<td>56.8</td>
<td>1.26</td>
<td>92.3</td>
<td>46.8</td>
</tr>
<tr>
<td>Head moves down starts brows move up</td>
<td>0.71</td>
<td>35.5</td>
<td>7.9</td>
<td>1.04</td>
<td>35.4</td>
<td>93.3</td>
</tr>
<tr>
<td>Mouth shape flat finishes brows move up</td>
<td>0.55</td>
<td>32.9</td>
<td>3.4</td>
<td>1.38</td>
<td>32.9</td>
<td>97.2</td>
</tr>
<tr>
<td>Brows move up before brows move down</td>
<td>0.52</td>
<td>0</td>
<td>10.8</td>
<td>$\infty$</td>
<td>100</td>
<td>18.5</td>
</tr>
<tr>
<td>Brows move up equals mouth shape flat</td>
<td>0.51</td>
<td>30.7</td>
<td>3.9</td>
<td>1.26</td>
<td>30.6</td>
<td>96.5</td>
</tr>
</tbody>
</table>

contrast, “brows move up before brows move down” does occur in other constructions, namely those in which a Topic or Hypothetical conditional clause (brows up) precedes a Wh-question.

“Head moves up starts brows move up” occurs in 33.5% of Yes/no questions but only 6.9% of other constructions. Half of the Yes/no questions are preceded by a topic; according to [1], two of the three possible topic markings involve head up. It is also claimed in [12] that head tilts forward with raised eyebrows in Yes/no questions. However, head behavior can also function parallel to body lean behavior, with tilt forward suggesting inclusion of the addressee and tilt back indicating exclusion of the addressee [167].

“Mouth shape flat finishes brows move up” occurs in 32.9% of the Yes/no questions as compared to only 3.4% of the other constructions, with a clear classification accuracy for the latter (97.2%). This is a truly surprising result which undoubtedly suggests further investigations in this direction as, to our knowledge, no function for flat mouth in ASL has been assigned in the existing literature. Since it spans the full duration of brows up (“brows move up equals mouth shape flat,” 30.7%) and ends when the brows up ends, these results suggest that this is a question mouth marker, although the issue is then raised as to why it is only not more frequent.
Table 2.7: Discriminant features for Assertions.

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>$\tilde{w}_i$</th>
<th>% Activation in Assertions</th>
<th>% Activation in Others</th>
<th>$d'$</th>
<th>% Classification in Assertions</th>
<th>% Classification in Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brows move down</td>
<td>1</td>
<td>17.1</td>
<td>46.2</td>
<td>0.85</td>
<td>82.9</td>
<td>55.9</td>
</tr>
<tr>
<td>Mouth shape round</td>
<td>0.61</td>
<td>39.9</td>
<td>61.9</td>
<td>0.56</td>
<td>62.6</td>
<td>63.3</td>
</tr>
<tr>
<td>Teeth close during brows move up</td>
<td>0.56</td>
<td>20.1</td>
<td>37</td>
<td>0.50</td>
<td>82.6</td>
<td>28.8</td>
</tr>
<tr>
<td>Brows move up</td>
<td>0.48</td>
<td>56.8</td>
<td>65.4</td>
<td>0.22</td>
<td>43.2</td>
<td>61.7</td>
</tr>
<tr>
<td>Mouth shape round during brows move down</td>
<td>0.46</td>
<td>3.2</td>
<td>19.6</td>
<td>0.99</td>
<td>96.8</td>
<td>27.6</td>
</tr>
</tbody>
</table>

**Assertions**

Assertions have been traditionally viewed as not marked by specific nonmanuials, leaving the articulators free to reflect ones that accompany nonmanually marked lexical signs as well as to reflect the signer’s emotional status. The cues identified as distinctive in Table 2.7 are notable for their relative absence in Assertions as compared to the other constructions. With respect to “brows move up,” the occurrence in Assertions is most likely due to the presence of topics with raised brows [1] prior to the Assertion itself.

**2.4.2 Experiment 2: Polarity Discriminant Features**

The study of polarity follows the same procedure described above. The discriminant features selected by the LOSO approach are given in Tables 2.8-2.11. These are the results for each of the four classes with polarity, *i.e.*, Hypothetical conditionals, Wh-questions, Wh-questions postposed and Assertions.

Here, we also performed the statistical significant analysis described in Experiment 1 section. All our results were again statistically significant with: $p < 10^{-32}$ for Hypothetical conditionals, $p < 10^{-19}$ for Wh-questions, $p < 10^{-24}$ for Wh-questions postposed and $p < 10^{-36}$ for Assertions.

Let us look at each of these results in more detail.
Table 2.8: Discriminant features for polarity in Hypothetical conditionals.

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>( \bar{w}_i )</th>
<th>% Activation in Positives</th>
<th>% Activation in Negatives</th>
<th>( d' )</th>
<th>% Classification in Positives</th>
<th>% Classification in Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head turns left before head turns right</td>
<td>1</td>
<td>8.3</td>
<td>31.3</td>
<td>0.9</td>
<td>91.7</td>
<td>31.5</td>
</tr>
<tr>
<td>Head turns left meets head turns right</td>
<td>0.94</td>
<td>16.6</td>
<td>70.8</td>
<td>1.5</td>
<td>83.4</td>
<td>70.8</td>
</tr>
<tr>
<td>Head turns right before head turns left</td>
<td>0.59</td>
<td>7.9</td>
<td>27.8</td>
<td>0.8</td>
<td>92.1</td>
<td>27.8</td>
</tr>
<tr>
<td>Head turns right during brow move up</td>
<td>0.53</td>
<td>32.4</td>
<td>66.7</td>
<td>0.9</td>
<td>67.6</td>
<td>66.7</td>
</tr>
<tr>
<td>Head turns right overlaps mouth shape other</td>
<td>0.51</td>
<td>17.9</td>
<td>46.3</td>
<td>0.8</td>
<td>82.1</td>
<td>46.3</td>
</tr>
</tbody>
</table>

Hypothetical Conditionals

From Table 2.8, we see that all notable features for polarity in Hypothetical conditionals are associated with head turns and are more frequent in negatives than in positives. This is an expected finding as negatives are generally marked by headshakes in ASL [148, 111, 123] and many other sign languages [58, 173, 77]. As discussed earlier, “brows move up” is associated with Hypothetical conditionals, and the occurrence of “brows move up” with negative Hypothetical conditional head turns suggesting that both conditionality and negation can be distinctly shown simultaneously without interfering with each other [161].

When we dig into the details of the temporal behavior of head turns, we identify linguistic interactions that have not been available to impressionistic analysis so far. We believe this is an improvement our algorithm has made possible for sign language research. In this sense, the findings with the ordering and the relation of head turns alert us to two previously unrealized findings about negative polarity in ASL.

The first finding is that the defining relation for negative polarity is “a head turn meets the opposite head turn” which kinematically correlates to a fast paced headshake. That the defining relation is “meets” rather than a head turn preference on either side of the relation is proved when we compare Table 2.8 with Table 2.11. In Table 2.8, what gives us the fast paced headshake is the “head turns left meets head turns right” discriminant feature. On the other hand, what gives us the fast paced headshake in Table 2.11 is the “head turns right
meets head turns left” discriminant feature. The commonality turns out to be the abstract linguistic relation “a head turn meets the opposite head turn.” The kinematic realization of this abstract linguistic property is a fast paced headshake.

The second finding is that we can generalize that negation normally begins with “head turns right.” Because this does not always occur, we state the general nonmanual marking as “a head turn meets the opposite head turn.” There is a widespread linguistic assumption that Assertions are the most basic, simplest clause type, and this is where we see the negative headshake start with “head turns right.”

When we look at the combination of Hypothetical conditional and negation, we are no longer looking at the simplest situation. Instead, the conditional contains the negation as part of its clause, and we expect the conditional marking to begin before the negation marking. In the case of constructions discriminant features for Hypothetical conditional, we determined that “head turns left during brows move up” is the discriminant feature for conditionals. As we will see in discriminant features for polarity in assertions, the primary indicator of fast paced negative headshake “head turns right meets head turns left” in Assertions, the most basic clause, starts with head turn to the right. In Hypothetical conditionals, the “head turns left” dominates the negative, and the fast paced negative headshake is modified to start on the left, yielding “head turns left meets head turns right”, the most active nonmanual marker in negative Hypothetical conditions (Table 2.8).

In addition to these two findings we also need to note that the headshakes reflected by “head turns right/left before head turns left/right,” with a short pause between the two, rarely occur in positive Hypothetical conditionals (7.9% and 8.3%), leading to 91.7% and 92.1% classification accuracy from the single feature of pausing alone. This observation supports our contention that assimilated “head turns left” starts the marking of negation and fast paced meeting of “head turns right” continues the marking. Without the “head turns right” as the second half of the fast paced negative headshake in positive conditionals, there is no purpose to the brief pause that separates the fast paced headshake from the rest.
of the head turns. Therefore, brief pauses between head turns highlights the separation of the fast paced negative headshake from the rest of the headshakes in negative conditionals. There is no need for these pauses in positive conditionals as the only head turns present are related to conditionality.

Beyond the results above, our results further highlight the role of the mouth in nonmanuals. Note the frequency of “mouth shape other” (meaning, not round or flat) during (overlapping with) “head turns right” in a large number of negative Hypothetical conditionals.

As we noted above, the fast paced “head turns left meets head turns right” gives us a strong cue for differentiating negative polarity from positive polarity in the conditional sentences. When this result is evaluated with “head turns right overlaps mouth shape other,” we come up with the pattern in Figure 2.8 where “mouth shape other” overlaps with the second half of the headshake. This temporal relation gives us another interesting and novel finding in that “mouth shape other” temporally occurs after the onset of negation as marked by the first head turn to left. Although the involvement of the mouth for negation in ASL had been detected in previous research [148], given the technology of the time, back then it was only possible to report the timing relation between the headshake and the hand movement, but not the exact temporal relation between the two nonmanual markers headshake and mouth position.

Next, note that the percentage of “mouth shape other” (46.3% vs. 17.9%) is strong enough not to be associated with a combined effect of lexical mouth-shapes of random signs.
in negative sentences. The contrast in discriminant percentages indicates that the mouth is actively involved in the expression of negation in ASL. This finding (82.1% and 46.3%, respectively) is consistent with results reported in [148]. In their study, negative sentences were compared to positive controls; headshakes were not always present in negatives, and whether or not headshakes were present, there was involvement of the mouth and/or chin in 96.5% of the negative productions. Furthermore, they noted that the most frequent combinations of nonmanual markings for negatives involved eyes (squished or closed) and a mouth position (corners of mouth down, mouth stretched, mouth tightly closed, chin contracted). These mouth positions are included in our coding of “mouth shape other.”

The fact that “mouth shape other” is not as frequent as headshake is another interesting finding. There are two ways to interpret this finding. First, although “mouth shape other” is present for almost half of the negative sentences, it could be a redundant or secondary prosodic cue, similar to the findings in [24] where the non-dominant hand is considered a secondary cue with respect to the primary cue of change in the mouth area tension. Therefore, “mouth shape other” would not need to occur as frequently as headshake. In this sense, headshake alone would be a sufficient prosodic cue for introducing negative polarity in conditionals. Second, the presence of “mouth shape other” could be a primary cue parallel to headshake. However, the combined semantic effect of headshake and “mouth shape other” may be more emphatic than the headshake alone. Therefore, the combination would only occur in situations where emphasis needs to be cued while headshake is more persistently present as the primary negative cue. Both of these possibilities need to be tested. The first one may be tested through prosodic perception studies while the second possibility may be tested with a semantic interpretation study. The upshot of the contribution of the current study is that the algorithm used in this study makes it possible for us to voice these two possibilities due to the temporal and distributional accuracy that we attain.
Table 2.9: Discriminant features for polarity in Wh-questions.

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>$\tilde{w}_i$</th>
<th>% Activation in Positives</th>
<th>% Activation in Negatives</th>
<th>$d'$</th>
<th>% Classification in Positives</th>
<th>% Classification in Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth closed meets teeth open</td>
<td>1.00</td>
<td>9.8</td>
<td>30.8</td>
<td>0.8</td>
<td>90.2</td>
<td>30.8</td>
</tr>
<tr>
<td>Teeth touch lip</td>
<td>0.89</td>
<td>25.2</td>
<td>6.7</td>
<td>0.8</td>
<td>25.2</td>
<td>93.3</td>
</tr>
<tr>
<td>Mouth shape other before head moves down</td>
<td>0.85</td>
<td>26.8</td>
<td>13.5</td>
<td>0.5</td>
<td>26.8</td>
<td>86.5</td>
</tr>
<tr>
<td>Mouth open starts brows move down</td>
<td>0.82</td>
<td>37.0</td>
<td>26.0</td>
<td>0.3</td>
<td>37.0</td>
<td>74.0</td>
</tr>
<tr>
<td>Teeth open during brows move down</td>
<td>0.82</td>
<td>39.8</td>
<td>63.5</td>
<td>0.6</td>
<td>60.2</td>
<td>63.5</td>
</tr>
<tr>
<td>Blink</td>
<td>0.76</td>
<td>69.5</td>
<td>76.9</td>
<td>0.2</td>
<td>76.8</td>
<td>46.2</td>
</tr>
<tr>
<td>Brows move down starts mouth open</td>
<td>0.74</td>
<td>8.1</td>
<td>15.4</td>
<td>0.4</td>
<td>91.9</td>
<td>15.4</td>
</tr>
<tr>
<td>Teeth open overlaps head turns right</td>
<td>0.74</td>
<td>11.0</td>
<td>30.8</td>
<td>0.7</td>
<td>89.0</td>
<td>30.8</td>
</tr>
</tbody>
</table>

Wh-questions

The discriminant features that distinguish negative and positive polarity in Wh-questions are more varied than those of Hypothetical conditionals and seem to be less clearly reflective of general negative marking. That is, they generally do not indicate head turns. Instead, a number of the features relate to mouth and teeth positions. In addition, there is no clear pattern of occurrence such as that seen with Hypothetical conditionals, where strong markings were seen for negatives as compared to positives. Here, sometimes a mouth or teeth feature is more prevalent in negatives and sometimes the reverse is true. This suggests that while Wh-questions can be clearly marked by brows down, when Wh-questions are negative, nonmanuals alone may not be able to carry both semantic functions. Such a conclusion is in keeping with two other observations in the literature. One is that both Wh-marking and negation use headshakes; the negative headshake is somewhat larger and slower [157]. The other is that whereas Yes/no questions rarely are marked by a manual sign and rely primarily on the brows up nonmanual marking, Wh-questions are most frequently accompanied by a manual Wh-sign. There are some notable examples where a Wh-question can occur without a Wh-sign, for example MANY “how many,” COLOR “what color” [93]. But reliance on Wh-signs means that nonmanuals may not be systematically recruited to carry the full load of semantic marking by themselves. These results suggest that when
negative and Wh-questions interact, nonmanuals like the mouth become more important.

Moreover, the current results define several interesting interactions in Wh-questions and polarity. “Brows move down starts mouth open” is highly classificatory (91.9%) for positive Wh-questions by its absence. While “brows move down” is clearly related to Wh-questions, mouth open could be related to some of the Wh-words being mouthed (e.g., who, what, which, when, why, etc.) and given the higher occurrence in negatives, possibly also ‘not.’ Another mouth cue with a high classification value (89%) for positives by its absence, “teeth open overlaps head turns right,” is almost three times more prevalent in negatives than in positives. Similarly, “mouth closed meets teeth open” is twice as prevalent in negatives as in positives and has a high classification value – its absence from positives yields correct classification 90.2% of the time despite its rare occurrence in Wh-questions in general. When such negative evidence (9.8%) is combined with positive evidence (30.8%), we may thus suggest that “mouth closed meets teeth open” is a candidate to discriminate between negative and positive polarity in Wh-questions. The computational model of this interaction is given in Figure 2.9.

Figure 2.9: Computational model of positive versus negative polarity in Wh-questions.

As we have discussed in the section above regarding negative conditionals, there is evidence of “mouth closed” as a marker of negation. The fact that it meets “teeth open” 30.8% of the time suggests that this cue may be interrupted by some lexical interference (mouthing of English words) tucked into the flow of prosody due to certain lexical items.

Another mouth feature that has a high classification value for negative Wh-questions
is “teeth touch lip,” which occurs in 25.2% of positives versus only 6.7% of negatives. This is likely the result of three of the positive Wh-questions containing signs that can be accompanied by mouthing of English words beginning with ‘f’ (fish, forks, finish).

**Wh-questions postposed**

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>( \bar{w}_i )</th>
<th>% Activation in Positives</th>
<th>% Activation in Negatives</th>
<th>( d' )</th>
<th>% Classification in Positives</th>
<th>% Classification in Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth shape other before mouth shape other</td>
<td>1</td>
<td>77.2</td>
<td>96.9</td>
<td>1.1</td>
<td>94.6</td>
<td>75</td>
</tr>
<tr>
<td>Mouth open</td>
<td>0.87</td>
<td>37</td>
<td>75</td>
<td>1.0</td>
<td>91.3</td>
<td>53.1</td>
</tr>
<tr>
<td>Head turns left during brows move down</td>
<td>0.87</td>
<td>46.7</td>
<td>96.9</td>
<td>1.9</td>
<td>94.6</td>
<td>65.6</td>
</tr>
<tr>
<td>Mouth open before mouth open</td>
<td>0.84</td>
<td>63</td>
<td>90.6</td>
<td>1.0</td>
<td>89.1</td>
<td>71.9</td>
</tr>
<tr>
<td>Head moves down during brows move down</td>
<td>0.84</td>
<td>45.7</td>
<td>75</td>
<td>0.8</td>
<td>54.3</td>
<td>75</td>
</tr>
<tr>
<td>Head turns left before mouth closed</td>
<td>0.83</td>
<td>28.3</td>
<td>71.9</td>
<td>1.2</td>
<td>90.2</td>
<td>43.8</td>
</tr>
<tr>
<td>Head turns left during mouth shape flat</td>
<td>0.78</td>
<td>3.3</td>
<td>37.5</td>
<td>1.5</td>
<td>96.7</td>
<td>37.5</td>
</tr>
<tr>
<td>Head tilts right during brows move down</td>
<td>0.77</td>
<td>13</td>
<td>59.4</td>
<td>1.4</td>
<td>87</td>
<td>59.4</td>
</tr>
<tr>
<td>Head turns left overlaps teeth open</td>
<td>0.75</td>
<td>15.2</td>
<td>53.1</td>
<td>1.1</td>
<td>84.8</td>
<td>53.1</td>
</tr>
</tbody>
</table>

In contrast to regular Wh-questions, there is a clearer pattern to negative marking for Wh-questions postposed, with discriminant features all occurring more frequently in the negatives than in the positives. This pattern seems to support the argument above concerning regular Wh-questions. The basic difference between Wh-questions with and without Wh-sign postposing is that when the Wh-sign occurs at the end of the question, the material that occurs before the Wh-sign does not have to be covered by Wh-marking. As discussed in discriminant features for polarity in Wh-questions, the material prior to the Wh-sign can sometimes be considered an Assertion, meaning that Wh-marking and negation marking would not come into conflict. Hence, Table 2.10 reflects features of negation on non-Wh-marked signs. This means that the nonmanuals can carry negation clearly, as seen by the prevalence of head turns among the discriminant features. This suggests a fundamental
linguistic difference between Wh-questions and Wh-questions postposed which confirms previous research [3].

Like regular Wh-questions, we see increased prominence of mouth and teeth positions which will require further research to explain, such as the interaction of mouth gestures with mouthing the English words when certain signs are produced [110]. Once again, this is an important, novel finding, reinforcing the previously overlooked suggestion of a more relevant mouth role in polarity [148].

**Assertions**

<table>
<thead>
<tr>
<th>ATL relation</th>
<th>( \bar{w}_i )</th>
<th>% Activation in Positives</th>
<th>% Activation in Negatives</th>
<th>( d' )</th>
<th>% Classification in Positives</th>
<th>% Classification in Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head turns left before head turns right</td>
<td>1</td>
<td>8.6</td>
<td>44.7</td>
<td>1.2</td>
<td>91.4</td>
<td>44.7</td>
</tr>
<tr>
<td>Head turns right before head turns left</td>
<td>0.97</td>
<td>10</td>
<td>43.6</td>
<td>1.1</td>
<td>90</td>
<td>43.6</td>
</tr>
<tr>
<td>Head turns right overlaps head turns left</td>
<td>0.97</td>
<td>62.4</td>
<td>97.2</td>
<td>1.6</td>
<td>96</td>
<td>68.9</td>
</tr>
<tr>
<td>Head turns left before head turns left</td>
<td>0.68</td>
<td>4.8</td>
<td>21.9</td>
<td>0.9</td>
<td>95.2</td>
<td>21.9</td>
</tr>
<tr>
<td>Head turns left before head turns right</td>
<td>0.61</td>
<td>71.1</td>
<td>97.8</td>
<td>1.5</td>
<td>96</td>
<td>69.7</td>
</tr>
<tr>
<td>Brows move up before head turns right</td>
<td>0.52</td>
<td>1.6</td>
<td>20.8</td>
<td>1.3</td>
<td>98.4</td>
<td>20.8</td>
</tr>
<tr>
<td>Head turns right meets head turns left</td>
<td>0.51</td>
<td>13.9</td>
<td>69.7</td>
<td>1.6</td>
<td>98.4</td>
<td>48.6</td>
</tr>
</tbody>
</table>

The results for polarity marking of Assertions also show clear nonmanual marking of negation, as all discriminant features occur more often in negatives than in positives. The primary cue in all discriminant features is head turn, reflecting negative headshakes. The only other discriminant cue is “brows move up,” which occurs before “head turns right;” this is the result of those Assertions that begin with a topic or are preceded by a conditional clause, both of which are marked with brows up, followed by a negative Assertion marked with headshake.

In sum, with the exception of Wh-questions, the marking of negative polarity is clear on the constructions included in this study, and Wh-questions themselves are known to differ from the other constructions in needing a manual Wh-sign most of the time. The surprises
in the data are related to mouth and teeth positions, which seem to gain prominence as nonmanual marking becomes more complex when multiple semantic functions are expressed simultaneously.

2.5 Discussion

Uncovering the discriminant features of the linguistic model governing nonmanuals in sign languages has proven to be an extremely hard problem. The present chapter shows how this can be resolved using a linguistic-computational approach. In this approach a linguistic representation of the face is first obtained. A computational approach is then employed to determine the combination of these features consistently observed in each class but not with others. The resulting linguistic model proves to be able to discriminate between nine different classes of sentences – Hypothetical conditionals, Wh-questions, Wh-questions postposed and Assertions in their two polarities and Yes/no questions in positive polarity.

The analyses described above strongly suggest that there are discriminant features that can be used to separate conditionals from non-conditionals, Yes/no questions from non-Yes/no-questions, Wh-questions and Wh-questions with postposed Wh-signs from non-Wh-questions, and Assertions from non-Assertions. In addition, for each of these except Yes/no questions which do not form negative in ASL, the discriminant features separate the negative structures from their positive counterparts. From the model (Table 2.3-2.11), the results indicate that some features are more relevant to accomplishing these distinctions than others. For example, blinks do not play a role in making these structural distinctions, nor was it expected that they would, as their function is more closely related to the marking of constituent structure (syntactic phrases) and the intonational phrasing that surrounds them [160]. Similarly, head tilts and head movements up and down appear to play no major role, leaving open the question of what their functions might be. Clearly the relevant features identified by these analyses are the head turns, brow positions, and mouth and teeth
features. The results for brow position confirm our expectations, both for “brows move up” and “brows move down.”

In addition, the algorithm gives temporal relations that are striking with respect to head turns, where there are both expected and important novel results. The use of multiple head turns as headshakes has been well-documented for ASL and other sign languages as a major nonmanual marker of negation [148, 111, 123, 58, 173, 77]. However, our findings with respect to temporal relations need to be emphasized because, as mentioned, although we know what makes a headshake, until now we did not have the means to quantitatively measure the temporal make-up of the interaction of the components of a headshake. In other words, the results of the present study suggest that not all temporal sequences of head turns left/right plus head turns right/left are the same. In negative conditionals and Assertions, negative polarity is most strongly cued when these meet one another, i.e., a faster paced headshake.

This opens a new venue for the study of headshakes. For instance, with regard to negative head turns it will be important to determine whether all negative headshakes are faster under all conditions across multiple sign languages. This possibility is raised in observations on Austrian Sign Language [86] concerning faster headshakes on negatives that follow regular speed headshakes on conditionals. The analysis can also be expanded to investigate if there are quantitative differences between languages that use headshake as a primary nonmanual cue such as ASL as compared to those that use negation as a secondary cue in addition to a different major nonmanual marker, such as Turkish Sign Language [60]. We also expect that these two novel findings for ASL will urge researchers of other sign languages to quantitatively investigate the nature of headshake since the surface cue, i.e., headshake, may very well be instantiated in more than one articulatory combination given the left and right directions of articulation, as well as temporal possibilities; a priori there is no reason to expect other sign languages which employ “headshake” as the major nonmanual cue to behave the same way as ASL does. On the big picture, this path also
opens up an exciting agenda, both for ASL and cross-linguistic research, for quantitatively detecting nuances in the behavior of certain nonmanual markers which look the same on the surface even to the eye of an experienced sign language annotator.

In addition to the insights about negation reported here, the approach presented in the present work also revealed that head turn left is a discriminant feature in conditionals. Again, work on Austrian Sign Language [86] is relevant for furthering research on this finding, noting that signers who are talking about things they think or wonder about use a higher, right side space. Conditionals are just such a possibility, as they indicate not fact but possibility, a hypothetical thought, possibly placed on the right for Austrian Sign Language. Comparing our findings with these in [86] opens up a research domain for further investigating crosslinguistic similarities and differences with the use space for conditionals.

Lastly, the results also highlight the important role that the mouth and teeth play in negation. It is noted in [148] that the most frequent combinations of nonmanual markings for negatives involved eyes (squished or closed) and a mouth position (corners of mouth down, mouth stretched, mouth tightly closed, chin contracted). These mouth positions are included in our coding of “mouth shape other,” which shows up as a discriminant feature overlapping with head turns in negatives. As we discuss above, the involvement of the mouth and teeth suggests importance of investigations in wider linguistic context to tease apart the possible secondary cue of “mouth shape other” from a possible interpretation of it as having a primary but emphatic function. Thus, these findings allow us to set up future studies by identifying the relevant variables that need to be controlled.

As a final note, it should be noted that the methodology described herein (and the implementation of the computational approach in Elan) will most probably find applications beyond the studies of sign language. Elan is a generic tool used in several disciplines and the statistical analysis described in the present chapter is equally valid in these studies.

43
2.6 Sentences Signed in our Database

Tables 2.12-2.13 list the simple sentences used in the present study. These correspond to four clauses defined in Section 2.1 (i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions). We also include the polarity of the sentences.

Tables 2.14-2.15 list the complex sentences used in the present study. These complex sentences are composed of a Hypothetical conditional clause that can be either positive or negative and a dependent clause that is either positive or negative, i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.

Table 2.16 lists the sentences signed by each one of the fifteen participants in our database. In order to obtain a larger variety of signed sentences, subjects signed different subsets. Also, the data used in the present paper corresponds to a total of 15 signers chosen from a larger set of 25 signers who originally participated in the data collection described in [66]. An experienced deaf researcher selected the 15 participants with highest clarity of performance and most fluency in ASL.
<table>
<thead>
<tr>
<th>Clause</th>
<th>Polarity</th>
<th>Construction type</th>
</tr>
</thead>
<tbody>
<tr>
<td># SARAH-i IX-3-i HAVE PARTY</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td># BRAD-i IX-3-i COOK FISH ON # GRILL IX-3-i</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td># BRAD-i IX-3-i NOT COOK</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># BRAD-i IX-3-i COOK FISH ON # GRILL NOT</td>
<td>Neg-postposed</td>
<td>Assertion</td>
</tr>
<tr>
<td># BRAD-i IX-3-i NOT COOK FISH ON # GRILL</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># JOHN-a IX-3-a a-BRING CHOCOLATE # CAKE WITH # NUTS</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td># BRAD-i IX-3-i COOK</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td># JOHN-a IX-3-a a-BRING</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td># JOHN-i IX-3-i FULL</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>CAN'T EAT FINISH EAT # CAKE</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># JOHN NOT BRING</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># JOHN-a IX-3-a NOT a-BRING CHOCOLATE # CAKE WITH # NUTS</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># JOHN CAN'T FINISH EAT</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># MARY-i IX-3-i NOT-YET START # GRILL</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># MARY START</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td># MARY-i IX-3-i START # GRILL</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td>FORK NOT KITCHEN</td>
<td>Neg-postposed</td>
<td>Assertion</td>
</tr>
<tr>
<td>FORK IN KITCHEN NOT</td>
<td>Neg-postposed</td>
<td>Assertion</td>
</tr>
<tr>
<td># SARAH-i IX-3-i HAVE PARTY Q++</td>
<td>Positive</td>
<td>Yes/no question</td>
</tr>
<tr>
<td># JOHN-i IX-3-i CAN EAT</td>
<td>Positive</td>
<td>Yes/no question</td>
</tr>
<tr>
<td># JOHN-a IX-3-a a-BRING CHOCOLATE # CAKE WITH # NUTS Q++</td>
<td>Positive</td>
<td>Yes/no question</td>
</tr>
<tr>
<td># MARY-i IX-3-i START Q</td>
<td>Positive</td>
<td>Yes/no question</td>
</tr>
<tr>
<td># MARY FINISH START</td>
<td>Positive</td>
<td>Yes/no question</td>
</tr>
<tr>
<td>FORK IX-4-a KITCHEN-a</td>
<td>Positive</td>
<td>Yes/no question</td>
</tr>
<tr>
<td>WHO COOK FISH ON # GRILL WHO</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>WHO COOK</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>WHO CAN EAT FINISH</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>WHO CAN'T EAT</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>WHY # MARY-i IX-3-i NOT START # GRILL WHY</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>WHO BRING</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td># SARAH POSS-3 PARTY WHEN</td>
<td>Positive</td>
<td>Wh-question postposed</td>
</tr>
<tr>
<td># JOHN-a IX-3-a NOT a-BRING CHOCOLATE # CAKE WITH # NUTS WHY</td>
<td>Negative</td>
<td>Wh-question postposed</td>
</tr>
<tr>
<td>IX-3-i # GRILL-i START WHO</td>
<td>Positive</td>
<td>Wh-question postposed</td>
</tr>
<tr>
<td>FORK WHERE</td>
<td>Positive</td>
<td>Wh-question postposed</td>
</tr>
</tbody>
</table>

Table 2.12: First set of sentences. The first column shows the sentence signed by a subset of the signers as described in Table 2.16. The second column indicates the polarity of the sentence i.e., positive or negative. The last column indicates the type of sentence i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.
Table 2.13: Second set of sentences. The first column shows the sentences signed by a subset of the signers as described in Table 2.16. The second column indicates the polarity of the sentence i.e., positive or negative. The last column indicates the type of sentence i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.
Table 2.14: First set of complex sentences. The first column shows the sentence signed by a subset of the subjects as described in Table 2.16. The second column indicates the polarity of the Hypothetical conditional clause and the polarity of the dependent clause i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.
<table>
<thead>
<tr>
<th>Clause</th>
<th>Complex sentences: Polarity and construction type of each of its parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF IX-2 NOTHING DO TOMORROW, # SARAH HAVE TAKE-UP PARTY</td>
<td>Negative Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>IF # JOHN-a IX-3-a NOT a-BRING CHOCOLATE # CAKE WITH NUT SPRINKLE, IX-1</td>
<td>Negative Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>WILL BRING IX-1 IF FORK NOT IX-3-a IN KITCHEN-a, IX-1 WILL GO GET</td>
<td>Negative Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>IF # JOHN-a IX-3-a NOT a-BRING CHOCOLATE # CAKE WITH NUT, WHO WILL BRING</td>
<td>Negative Hypothetical conditional+Positive Wh-questions</td>
</tr>
<tr>
<td>WHO well IF # JOHN-I IX-3-I CAN'T RUN-OUT EAT THAT CAKE, WHO CAN</td>
<td>Negative Hypothetical conditional+Positive Wh-questions</td>
</tr>
<tr>
<td>IF # SARAH HAVE PARTY TOMORROW, IX-1 CAN'T GO</td>
<td>Positive Hypothetical conditional+Negative Assertion</td>
</tr>
<tr>
<td>IF FORK HAVE IX-3-a KITCHEN-a, IX-1 WONT GO GET IF # JOHN-1 IX-3-i</td>
<td>Positive Hypothetical conditional+Negative Assertion</td>
</tr>
<tr>
<td>RUN-OUT THAT CAKE, IX-1 NOT SURPRISE IX-1 IF # BRAD-I IX-3-i COOK</td>
<td>Positive Hypothetical conditional+Negative Assertion</td>
</tr>
<tr>
<td>FISH ON # GRILL, IX-1 WON'T EAT IX-1 IF IX-3-a FORK IN KITCHEN-a, WHY</td>
<td>Positive Hypothetical conditional+Negative Wh-questions</td>
</tr>
<tr>
<td>IX-1 CAN'T SEARCH-FIND WHY well IF IX-3-i MARY-I KNEW WILL RAIN, WHY-</td>
<td>Positive Hypothetical conditional+Negative Wh-questions</td>
</tr>
<tr>
<td>NOT IX-3-i PROCEED START GRILL well IF # SARAH HAVE PARTY TOMORROW,</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>IX-1 WILL BRING # CAKE IF # JOHN-a IX-3-a a-BRING CHOCOLATE # CAKE</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>WITH NUT SPRINKLE, IX-1 WILL BRING FORK IF FORK LEFT IX-3-a KITCHEN-a,</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>IX-1 WILL GO GET IF # MARY PROCEED START # GRILL, IX-1 WILL PROCEED</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>COOK FISH IF # BRAD-I IX-3-i COOK FISH ON # GRILL, IX-1 WILL</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>PROCEED BRING SALAD IF # BRAD-I IX-3-i PROCEED COOK FISH ON # GRILL,</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>WILL EXPLODE WILL IF IX-1 KNEW THAT IX-3-i # BRAD-I PROCEED COOK</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>FISH ON # GRILL, IX-1 PREFER ORDER # PIZZA IF # SARAH HAVE PARTY</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td>TOMORROW, WHO BRING FOOD WHO well IF # MARY-I IX-3-i PROCEED START</td>
<td>Positive Hypothetical conditional+Positive Assertion</td>
</tr>
<tr>
<td># GRILL, WHO PROCEED COOK FISH WHO well IF FORK IX-3-a IN KITCHEN-a,</td>
<td>Positive Hypothetical conditional+Positive Yes/no questions</td>
</tr>
<tr>
<td>IX-3 IN DRAWER IN Q+i+ IF # MARY-I IX-3-i PROCEED START # GRILL, MIND</td>
<td>Positive Hypothetical conditional+Positive Yes/no questions</td>
</tr>
<tr>
<td>IX-2 COOK FISH MIND IX-2 IF # JOHN-I IX-3-i PROCEED RUN-OUT EAT THAT</td>
<td>Positive Hypothetical conditional+Positive Yes/no questions</td>
</tr>
<tr>
<td>CAKE, MIND IX-1 DRINK THAT LAST BEER IF IX-3-i # BRAD-I PROCEED COOK</td>
<td>Positive Hypothetical conditional+Positive Yes/no questions</td>
</tr>
<tr>
<td>FISH ON # GRILL, MIND IX-2 EAT IX-2</td>
<td>Positive Hypothetical conditional+Positive Yes/no questions</td>
</tr>
</tbody>
</table>

Table 2.15: Second set of complex sentences. The first column shows the sentence signed by a subset of the subjects as described in Table 2.16. The second column indicates the polarity of the Hypothetical conditional clause and the polarity of the dependent clause i.e., Yes/no questions, Wh-questions, Wh-questions postposed and Assertions.
Table 2.16: Sets of sentences signed by the chosen fifteen participants. The first column is the participant code. Every session was divided between two and four blocks. In each block the participant signed sentences belonging to the ones described in Tables 2.12-2.15.

<table>
<thead>
<tr>
<th>Signer Code</th>
<th>First Block</th>
<th>Second Block</th>
<th>Third Block</th>
<th>Forth Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Table 2.13</td>
<td>Table 2.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Table 2.13</td>
<td>Table 2.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.12</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
<td>Table 2.12</td>
</tr>
<tr>
<td>5</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
</tr>
<tr>
<td>6</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
</tr>
<tr>
<td>7</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
</tr>
<tr>
<td>8</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
</tr>
<tr>
<td>9</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
<td>Table 2.15</td>
</tr>
<tr>
<td>10</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
<td>Table 2.12</td>
</tr>
<tr>
<td>11</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td>Table 2.15</td>
<td>Table 2.12</td>
</tr>
<tr>
<td>12</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Table 2.12</td>
<td>Table 2.12</td>
<td>Table 2.14</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

The Not Face: A Grammaticalization of Facial Expressions of Emotion

3.1 Summary

Facial expressions of emotion are thought to have evolved from the development of facial muscles used in sensory regulation and later adapted to express moral judgment. Negative moral judgment includes the expressions of anger, disgust and contempt. Here, we study the hypothesis that these facial expressions of negative moral judgment have further evolved into a facial expression of negation regularly used as a grammatical marker in human language. Specifically, we show that people from different cultures expressing negation use the same facial muscles as those employed to express negative moral judgment. We then show that this nonverbal signal is used as a co-articulator in speech and that, in American Sign Language, it has been grammaticalized as a nonmanual marker. Furthermore, this facial expression of negation exhibits the theta oscillation (3-8 Hz) universally seen in syllable and mouthing production in speech and signing. These results provide evidence for the hypothesis that some components of human language have evolved from facial expressions of emotion, and suggest an evolutionary route for the emergence of grammatical markers.
3.2 Introduction

The use of facial expressions for nonverbal communication is well documented in primates [37, 144]. In humans, these facial expressions can be used as a non-verbal signal or as linguistic markers in language and thus as part of its grammar [11, 90, 85, 159, 124]. A longstanding question in science is: where do these linguistic markers come from? Here, we study the hypothesis that at least some of these linguistic markers are grammaticalizations of facial expressions of emotion. That is, facial expressions of emotion have evolved to serve a grammatical function and may be precursors to the development of human language [19, 52, 98, 84]. In particular, we propose that facial expressions of negative moral judgment have evolved into a marker of negation in speech and signing. This previously unreported facial expression of negation is shown to use a unique combination of the muscle actions necessary to express negative moral judgment and be in the theta frequency band of speech and sign production.

Facial expressions of emotion are thought to have evolved through the development of facial articulations used in sensory regulation, e.g., closing the airways and eyes when expressing disgust to avoid disease [4, 29, 142, 87]. However, anger, disgust and contempt have been adapted to convey negative moral judgment, e.g., violations of ones rights, societal norms or beliefs [118, 64, 132, 138, 45]. Fig. 3.1 shows a sample image of each of these facial expressions. The production of each of these expressions is unique [46], meaning that the facial articulations (known as Action Units, AUs) used to produce these expressions are distinct from one another. Different facial articulations are identified with a distinct AU number. For example, anger is produced with AUs 4, 7 and 24, disgust with AUs 9, 10, 17, and contempt with AU 14 [44], Fig. 3.1.

As shown in Du et al. [42], when facial expressions of emotion are compounded to create new categories, the resulting expression is defined using a subset of the AUs employed by the subordinate categories. For example, the facial expression of angrily disgusted is produced
with AUs 4, 7, 10 and 17, a subset of those used to express its subordinate categories anger and disgust, which include AUs 4, 7, 9, 10, 17, 24. This new combination of AUs must be distinct from the ones employed to express any other emotion category; otherwise, the resulting category could not be visually differentiated from the rest. The combination of AUs used to express angrily disgusted, for instance, is distinct from those seen when expressing other emotions [42].

Since the AUs used to express negative moral judgment are 4, 7, 9, 10, 14, 17 and 24, a facial expression of negation that has evolved through the expression of negative moral judgment should be defined by a unique subset of these AUs. Indeed, we find that the facial expression of negation is constructed using such a unique subset. In particular, this expression of negation is found to employ AUs 4, 14, 17 and 24, Fig. 3.2.

Figure 3.2: A few examples of people conveying negation with a facial expression. People from different cultures use AUs 4, 17 and either 14 or 24 or both.
To test the hypothesis that this expression is used as a co-articulator in speech and as a nonmanual marker in signing, the expression defined in the preceding paragraph must be found in speech and signing of negative clauses across languages. We analyze a large number of video clips of English, Spanish and Mandarin Chinese speech as well as American Sign Language (ASL) and identify the same facial expression described above as a marker of negation in all languages (spoken and signed).

Additionally, the hypothesis that facial expressions of emotion evolved to have grammatical function is supported by the observation that both speech and facial articulations share the same frequency of production, between 3 and 8 Hz (theta oscillation) [28, 98]. This is true even in facial expressions used by primates, e.g., lip-smacking [59]. Increasing or decreasing these frequencies beyond these margins reduces intelligibility of speech and facial articulations. This is thought to be due to oscillations in auditory cortex, which allows us to segment theta-band signals from non-coding distractors [65]. Thus, if facial expressions of emotion evolved to become linguistic markers, these should also be produced within this intelligible theta rhythm. Indeed, we find that the production of the identified facial expression of negation exhibits this rhythm in all languages (spoken and signed).

In sum, to the authors knowledge, these results provide the first evidence for the grammaticalization of the facial expressions of negative moral judgment as a linguistic marker of negation, and suggest a possible evolutionary route from expression of emotion to human language.

### 3.3 Results

We study the hypothesis that a linguistic marker of negation has evolved through modifications of the facial expressions of negative moral judgment. We provide evidence favoring this hypothesis by showing that this facial expression of negation is universally produced by people of different cultural backgrounds and, hence, defined by a unique subset of AUs
Table 3.1: Consistently used Action Units (AUs) by subjects of different cultures when producing a facial expression of negation.

<table>
<thead>
<tr>
<th>AU #</th>
<th>Percentage of people using this AU</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>96.21</td>
</tr>
<tr>
<td>17</td>
<td>71.21</td>
</tr>
<tr>
<td>14 or 24</td>
<td>71.97</td>
</tr>
</tbody>
</table>

originally employed to express negative moral judgment (Experiment 1), used as a co-articulator in speech (Experiment 2), employed as a linguistic nonmanual marker in signing (Experiment 3), and produced at an intelligible theta rhythm (Experiment 4).

### 3.3.1 Experiment 1. Nonverbal Expression of Negation

Participants were asked to produce a facial expression that clearly indicated negation, *e.g.*, negating a statement. No instructions on which muscles to move were given. Participants had the opportunity to practice the expression in front of a mirror before picture acquisition. Pictures of the facial expression were taken at the apex (*i.e.*, maximum deformation of the face). A total of 158 subjects of different cultural and ethnic backgrounds participated in this experiment. The mother tongue of the participants thus varied. All pictures were frontal view. Images were manually coded using FACS (Facial Action Coding System) [44] to identify the AUs used by each participant. Table 3.1 shows the percentage of AUs in these images. Participants consistently used AUs 4, 17 and either 14 or 24 or both, Fig. 3.2. AU 7 was sometimes added but by a significantly smaller percentage of subjects (37.88%). The consistency of use of AUs 4, 17 and either 14 or 24 is comparable to the consistency observed in AU usage in universal facial expressions of emotion (>70%) [42].

As can be seen in Table 3.1 and Fig. 3.2, the AUs used to communicate negation correspond to a subset of those employed to express negative moral judgment. AUs 4 comes from the facial expression of anger, AUs 17 and 24 from disgust, and AU 14 from contempt, Fig. 3.3. AU 4 lowers the inner corner of the eyebrows, AU 17 raises the chin, and AUs 14 and 24 (which are rarely co-articulated) press the upper and lower lips against...
one another. Since the actions of AU 14 and 17 results in pressing the lower and upper lip against one another, only one of them is necessary. AU 7 is only used by a small number of people. Its role is to tighten the eyelids and is commonly used to express anger. We refer to this facial expression of negation as the “not face.”

Figure 3.3: We test the hypothesis that compounding the expressions of negative moral judgment (anger, disgust, contempt) yields a grammaticalized universal facial expression of negation. Shown on top are the three facial expressions of anger, disgust and contempt and their corresponding AUs. The AUs with an outer box are the ones that are combined to express negation (bottom image).

The consistency of AU production of the “not face” across participants of different cultural backgrounds suggests that this facial expression of negation is universally used across populations. To see if this finding generalizes beyond images collected in the lab, we used a computer vision algorithm (Materials and methods) to find spontaneous samples of the “not face” in the web. Specifically, we automatically analyzed images and videos of sports. We selected sports because players usually negate the action of the referee (e.g., after a foul) or themselves (e.g., after missing an open shot) using non-verbal gestures. Thus, it is expected that some players will use the above-identified facial expression of negation. Additionally,
these expressions are spontaneous, which will allow us to determine if the AU pattern seen in lab conditions is consistent with that used by people in naturalistic environments. Our search identified several instances of the “not face,” Fig. 3.4, demonstrating that the “not face” defined above is also observed in spontaneous expressions of negation.

![Figure 3.4: Samples of spontaneous expressions of negation. 3.4(a)-3.4(b). Pictures of professional soccer players after missing an open shot. 3.4(c). A coach disagreeing with the referee.](image)

In summary, the results of this first experiment demonstrate the universality of this nonverbal gesture. Next, we study if this expression is used as a co-articulator in speech across languages. The concept of negation changes somewhat as it is brought from nonverbal gesture into language, as the meanings expand to include those possible in language but not outside of language. The relevance here is that nonverbal negation such as that elicited in Experiment 1, when subjects are asked to produce a face equivalent to negation in speech, is limited to meanings like denying permission (You can’t do that), responding to requests / offers (No, I don’t want that), agreements (No, I don’t want to), or information states (No, I don’t know the answer). Brought into language, negation also allows us to deny the truth of a statement (No, that’s not right he didn’t do that) and to provide the speaker’s evaluation of another’s actions (Unfortunately, he failed the test), among other more complex meanings.
3.3.2 Experiment 2. Co-articulation of the Not Face in Speech

We filmed speech production of English, Spanish and Mandarin Chinese of twenty-six participants. Speech production for a total of 78 sentences, corresponding to four clause types and the two polarities, were recorded. Specifically, nineteen of these sentences had negative polarity (see Supporting Information for a list of the sentences). The same sentences were collected in each language. Native speakers read one sentence at a time, memorized it, and then reproduced it from memory while looking frontally at the camera. Subjects were instructed to imagine they were having a conversation with the experimenter, who remained behind the camera for the entirety of the experiment. This allowed us to film more naturalistic speech. Filming was uninterrupted, meaning that subjects were filmed while reading, memorizing and uttering these sentences. Additionally, we collected subject responses to four questions that can yield a negative response (e.g., a study shows that tuition should increase 30% for in-state students, what do you think?; a list of these questions is in the Supporting Information). Subjects were filmed while responding to these questions.

All frames of these video sequences were FACS coded using a computer vision algorithm (Materials and methods). Frames with a facial expression produced with only AUs 4, 17 and either 14 or 24 or both were identified. This approach found 41 instances of the “not face.” Twelve of these instances were produced during a negative sentence, seventeen were used by the subjects to indicate to the experimenter that they had forgotten the sentence and wanted to read it again, ten followed the incorrect reproduction of a sentence, and two were during the production of positive sentences by two English speakers, Table 3.2. The twelve examples of co-articulation of the “not face” and speech in negative sentences are observed in all clauses and were found in English, Spanish and Mandarin Chinese. In contrast there were only two instances of the “not face” in positive sentences (a hypothetical and a yes/no-question), both produced by English speakers. This corresponds to 0.13% of the positive sentences in our dataset, compared to the production of the “not face” in 2.83% of the negative sentences. We note that the “not face” was more commonly produced by English
Table 3.2: Usage of the “not face” in different languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Negative Sentence</th>
<th>Positive Sentence</th>
<th>Incorrect Performance</th>
<th>Forgotten Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Spanish</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Mandarin</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>17</td>
</tr>
</tbody>
</table>

The second and third column indicate that the expression was produced during a negative or positive sentence (co-articulator), respectively. The fourth column shows the number of times that participants used the “not face” to indicate a wrong reproduction of the intended sentence (non-verbal signal). The fifth column indicates the number of times that participants forgot the memorized sentence; these cases were immediately followed by a request to re-read the sentence. No other instances of the not face were identified. Thus, the “not face” was used in 2.83% of the negative sentences, but in only 0.13% of the positive sentences.

speakers than Spanish or Mandarin Chinese speakers. It is well-known that Americans emphasize expressions more than subjects from other cultural backgrounds [43, 104]. A recent study [133] shows that expressivity is correlated with the cultural diversity of the source migrants, suggesting a mechanism for exaggerating expressions in countries like the US. This seems to be the case for the expression of negation.

These results demonstrate the co-articulation of the “not face” in speech in negative sentences. As before though, one may ask if these expressions are also found outside the lab. To study this, we used a computer vision algorithm (Materials and methods) to identify spontaneous samples of the “not face” in 52 minutes of videos of American television. We identified two instances where people used the “not face” in negative sentences, Fig. 3.5. No instance of the “not face” was found in positive sentences.

3.3.3 Experiment 3. Grammatical (nonmanual) Marker in Signing

In sign languages, some facial expressions serve as nonmanual markers and thus have grammatical function [11]. A linguistic nonmanual marker can co-occur with a manual sign of
the same grammatical function or can be the sole marker (without the manual component). Zeshan [174] separates sign languages that can negate a sentence using only a nonmanual marker (“nonmanual dominant”) from those that require a manual sign (“manual dominant”). A typical nonmanual marker of negation in ASL is a headshake, which changes the polarity of a sentence even in the absence of the manual sign for “not” [148, 174, 122].

Our hypothesis is that the “not face” is used as a nonmanual marker of negation in ASL. We used a computer vision algorithm (Materials and methods) to identify occurrences of the “not face” in videos of fifteen ASL signers signing the same 78 sentences of Experiment 2 (see Supporting Information for a list of these sentences). We identified 114 instances of the “not face” in the signing of negative sentences. This corresponds to 15.4% of the negative sentences we analyzed. The number of occurrences is comparable to the number of negative sentences that include the nonmanual marker headshake, which is 14.75% of the sentences in our database. Further, our analysis revealed that the “not face” was used to indicate negation: in conjunction with the manual sign “not,” in combination with the nonmanual marker headshake, and as the sole marker of negation (i.e., without a headshake or a manual marker of negation). This result thus demonstrates the “not face” performs grammatical function in ASL.

In summary, the results described in Experiments 1-3 support the existence of a facial expression of negation that has been grammaticalized through the modification of facial
expressions of emotion. However, a grammaticalized “not face” should also have a frequency of production similar to that seen in speech and signing, i.e., a theta oscillation (3-8 Hz). We test this prediction next.

3.3.4 Experiment 4. Frequency of Production

Syllable production and mouthing in speech universally exhibits a theta oscillation [28]. Similarly, ASL syllables are signed in the theta range [165]. Cells in auditory cortex also oscillate at this frequency which is thought to facilitate segmentation of the speech (theta-oscillation) signals from non-coding distractors oscillating at other frequencies [65]. If the facial expression of negation identified above has indeed been grammaticalized and is thus interpreted by the brain as a grammatical marker, its rhythmicity should parallel that seen in speech and signing.

First, we measured the frequency of production of the “not face” in the videos of English, Spanish and Mandarin Chinese collected in our lab. The frequency is defined as the number of frames per second over the number of consecutive frames with the target expression and is thus given in Hz (Materials and methods). The frequency of production of the “not face” was found to fall within the theta band, with the mean over all spoken languages at 5.68 Hz. This was also true for each of the languages independently. In English, the “not face” was produced at 4.33 Hz. Spanish speakers had a similar production frequency of 5.23 Hz, while Mandarin Chinese speakers had a slightly faster tempo of 7.49 Hz.

We repeated the above procedure to compute the frequency of production of the “not face” in ASL. That is, we analyzed the video sequences of ASL described in Experiment 3. As in speech, the mean frequency of production was identified at 5.48 Hz and, hence, falls within the theta band as well.
3.4 Discussion

We identified, for the first time, a facial expression of negation that is consistently used across cultures as a non-verbal gesture and in languages as a co-articulator and a grammatical marker. Crucially, our results supported a common origin of the “not face” in facial expressions of emotion, suggesting a possible evolutionary path for the development of non-verbal communication and the emergence of grammatical markers in human language.

3.4.1 Universality of Facial Expressions

Greek philosophers [9], Descartes [39], Hume [79] and Darwin [37], among others, have been interested in the universality of facial expressions. By universal, it is understood that these expressions are innate and, hence, are primarily of a biological origin, rather than culturally bound; i.e., the classical scientific question of nature versus nurture [82]. When the same production of an expression is found in a variety of cultures, this is taken as strong evidence of its universality [45]. This means that the AUs employed to generate that facial expression must be the same across cultures. Importantly, this identical production has to be elicited when people from different cultures wish to express the same concept / category. Furthermore, this production needs to be unique, i.e., this expression must be different from the production of expressions of other categories; otherwise it would not be visually distinctive from other percepts.

To date, most of the work to identify universal facial expressions has focused on the expression of emotion. Using this approach, twenty-one facial expressions of emotion have been identified [42]. These emotional expressions have been shown to be systematically produced using the same AU combinations by people of different cultures. These expressions have also been shown to have a unique AU combination and thus be visually distinctive from one another. For instance, when people of different cultures want to express anger, they use the same, unique combination of AUs.
If these facial expressions of emotion are indeed universal and, hence, have a common biological origin, the next question of interest is to understand how they came to be. Darwin hypothesized that some expressions evolved through sensory regulation or to create protective behavior. For example, when expressing anger, the brow furrow protects our eyes from punches. In this example, our capacity to produce furrowed brows is thought have evolved to protect our sensors. On the other hand, disgust enhances central vision and limits inhaling by closing breathing pathways, protecting us from germs. These expressions were later adapted to convey emotions and moral judgment to observers. For example, the same facial muscles are used to express moral disgust as gustatory distaste.

The present chapter identified, for the first time, a facial expression of negation that is produced using the same AUs by people of different cultures, suggesting this expression is universal. We then studied the origins of this expression. People express negative moral judgment (moral anger, disgust and contempt) when their rights, beliefs, and societal norms are violated. Since these expressions are used to indicate negative disagreement, we hypothesized that the facial expression of negation evolved through the expression of moral anger, disgust and contempt.

Previous research has shown that when a facial expression is a compound of several subordinate categories (as in our hypothesis), the AUs used to produce this expression are a subset of the AUs used by its subordinate categories. For example, angrily disgusted is produced with AUs 4, 7, 10 and 17, a subset of those used to express anger and disgust. Thus, if the “not face” evolved from the expression of moral anger, disgust and contempt, then its expression should be constructed using a subset of the AUs of these three expressions, Fig. 3.3. We found this to be the case, supporting the view that the “not face” originated in the expression of negative moral judgment.

Moreover, we found the exact same production in non-verbal gesturing (without speech) in people of distinct cultures, as a co-articulator in speech in three spoken languages, and as
a grammatical marker in ASL. The extensive use of this expression further supports the uni-
versality hypothesis and, crucially, illustrates a possible evolutionary path from emotional
expression to the development of a grammatical marker of polarity. To our knowledge,
the results reported herein are the first to demonstrate the evolution of grammatical facial
markers through the expression of emotion, illustrating a possible path for the emergence
of language.

3.4.2 The Rhythms of Facial Action Production in the “Not Face” and
Syllable Production in Language are Consistent

Communicative signals can be very difficult to analyze. To resolve this problem, it is believed
that the brain takes advantage of statistical regularities in the signal. For example, speech
production, and its associated facial articulations, universally follow a theta-based rhythm
[28]; this also holds true in signing [165]. Thus, to discriminate coding (i.e., important
communicative) signals from non-coding distractors, the brain only needs to process signals
that fall within the theta band (3 to 8 Hz). Supporting this theory, cells in auditory
cortex are known to oscillate within this range, which may facilitate the segmentation of
coding versus non-coding signals [65]. Furthermore, lipreading in hearing adults [27] and
signing in congenitally deaf adults [117, 115] activate auditory cortex. And, non-verbal
facial expressions in non-human primates are also produced within this theta range [59],
suggesting that verbal and non-verbal communication signals are identified by detecting
signals in this frequency band.

We found that the “not face” is also universally produced using this rhythm. This
was true in English, Spanish, Mandarin Chinese and ASL. Thus, this facial expression
would be identified as a coding segment of a non-verbal communicative signal. This result,
compounded with the universality of AU production, suggests that the “not face” is used by
humans to communicate and visually interpret negation and that this signal can be readily
incorporated as a grammatical marker in languages.
3.4.3 A Linguistic Nonmanual Marker of Negation

Previous studies have primarily focused on the grammaticalization of the headshake [122]. However, there is still debate on the role and origins of this linguistic marker. Kendon [83], for instance, reports that headshakes are widely, but not universally, used across cultures, with backwards head tilt being one frequent alternative. And, importantly, there is no convincing data on the origins of the headshake, with feeding avoidance in babies a putative unsubstantiated assumption [122].

Focus on the negative headshake and its status as a linguistic marker rather than merely a gesture has generally led to minimal attention paid to the negative facial expressions that occur with or without these headshakes in signing. The results reported in the present study have not only identified a nonmanual marker of negation, but also provided evidence of their possible evolution through the expression of emotion. This lays the foundation for a research program focusing more specifically on facial expressions both in the context of and in the absence of associated head movements and their origin on the expression of emotion.

For example, a sad expression usually accompanies the manual sign for sad [129], yet little is known about the role of this and other facial expressions in signing. One reason for the lack of studies of facial expressions in signing is the large number and complexity of observed facial articulations during signing. Nonetheless, with the advance of automatic analysis of facial expressions using computer vision algorithms, such studies are now feasible and starting to yield important novel findings [16]. Similarly, research in German Sign Language shows that facial expressions of emotion co-articulated with their manual signs might have changed over time, suggesting a grammaticalization of these expressions [47]. These studies reported differential rhythms and placement of these expressions within a sentence. The results reported in the present study suggest that the rhythmic nature of these adaptations are essential to justify a grammaticalization of these expressions and their grammatical function, an area of research that needs further attention.
3.4.4 From Expression of Emotion to Grammatical Function

Emotions dominate much of cognition [119]. This is evident, for example, in the integration of affect and reasoning, where thoughts are given an affective valance [64]. This association process, which may include reciprocity, guilt and embarrassment, are thought to be (in part) the result of natural selection [53]. Some of these emotions are externalized in the form of facial expressions. Negative moral judgment is one example of this system. Its expressions can be used to indicate anger, disgust or contempt on violations to ones rights, societal norms or beliefs [132, 138]. Recent research studies suggest that these facial expressions first appeared as a protective mechanism or as sensory regulations, e.g., to avoid contact with germs in disgust [4, 29, 142, 87]. Only later, did humans find them useful to communicate emotions non-verbally [37].

On the other hand, research in sign languages has identified facial expressions used as grammatical markers [122]. For instance, moving the eyebrows up (which involves AUs 1 and 2) is used as a marker of yes/no-questions. Unfortunately, we do not yet understand how these grammatical markers evolved. The studies in the present chapter illustrate how these grammatical markers could have emerged. Our results support the view that at least some of these grammatical markers have evolved through the expression of emotion.

The fact that some of these facial expressions have acquired grammatical function has implications for the understanding of how human language emerged [19, 52] and offers a unique opportunity to study the minimally required abstract properties of human linguistic knowledge [30]. The present study, for example, suggests a very recent evolution of (at least) the herein identified linguistic marker we have called the “not face.” However, expressions of emotion evolved over a much larger timespan, providing a plausible explanation for the apparent sudden development [21, 75] of grammatical markers through the grammaticalization of facial expressions of emotion.
3.5 Materials and Methods

3.5.1 Participants

The experiment design and protocol was approved by the Office of Responsible Research Practices at The Ohio State University (OSU). A total of 184 human subjects (94 females; mean age 23; SD 5.8) were recruited from the local community and received a small compensation in exchange for participation. In addition, we analyzed the ASL dataset of Benitez-Quiroz et al. [16], which includes 15 native or near-native signers of ASL.

3.5.2 Data Acquisition

Subjects participating in Experiments 1-2 were seated 1.2 meters from a Canon IXUS 110 camera and faced it frontally. Two 500-W photography hot lights were located left and right from the midline, passing through the center of the subject and the camera. The light was diffused with two inverted umbrellas, i.e., the lights pointed away from the subject toward the center of the photography umbrellas, resulting in a diffuse lighting environment.

In Experiment 1, participants were asked to produce a facial expression of negation. Additionally, the experimenter suggested possible situations that may cause a negative answer and expression. Then, participants were photographed making their own facial expression of negation. Crucially, subjects were not asked to move specific muscles or facial components. Instead, they were encouraged to express negation as clearly as possible while minimizing head movement (i.e., expressing negation without a headshake). Photographs were taken at the apex of the expression and were full-color images of 4000 × 3000 pixels.

In addition, a neutral facial expression of each subject was collected. A neutral expression is one where all the muscles of the face are relaxed (not active) and thus no AU is present.

In Experiment 2, a computer screen was placed to the right of the camera to allow subjects to read and memorize a set of sentences in their native language. Specifically, a sentence was shown on the screen and participants were asked to memorize it. The
experimenter instructed participants to indicate when they had memorized this sentence, at which point the screen went blank. Then, participants were asked to reproduce the memorized sentence as if speaking (frontally) to the experimenter, who remained behind the camera for the entirety of the experiment. After successful completion of this task, a new sentence was displayed on the computer monitor and the same procedure was repeated. Data acquisition of the sentences in Experiment 2 took an average of 20 minutes.

The data used in Experiment 3 is that collected by the authors in a previous study [16]. Importantly, the sentences used in Experiments 2 and 3 were the same. A list of these sentences can be found in the Supporting Information.

Data acquisition was always completed using the participant’s native language (English, Spanish, Mandarin Chinese or ASL).

3.5.3 Manual Analysis of the Facial Expressions in Experiment 1

An AU is the fundamental action of individual or group of muscles that produces a visible change in the image of a face. For example, AU 4 is defined as the contraction of two muscles that in isolation produces the lowering of the eyebrows. To detect this movement in each of the images of Experiment 1, a FACS expert (coder) compared the images of the collected “not face” with an image of a neutral expression of the same subject. This allowed the FACS coder to successfully identify those AUs present in the image.

We manually FACS coded the data collected in Experiment 1. We identified AUs that were present in more than 70% of the subjects, since this is known to show universal consistency across subjects of distinct cultural upbringings [42]. This identified the AUs in Table 3.1. We then identified AUs with >30% consistency (i.e., AUs used by a small number of participants only). Only one such secondary AU (AU 7) was found to be used by of the participants (see Supplemental Excel file with detailed AU coding).
3.5.4 Computer Vision Approach for AU Detection

In Experiments 1 and 2, we FACS coded images and videos downloaded from the internet. We used Google® Images with different combinations of search keywords: disagreement, negation, soccer, referee, missed penalty. We also used videos downloaded from Youtube® using the same combinations of keywords. The number of images to be FACS coded was too large to be completed manually. Thus, we used a commercial software (Emotient 5.2.1, http://emotient.com) to identify possible candidate frames with the AUs that define the “not face.” This software uses state-of-the-art computer vision algorithms to identify AUs in images and video. To improve the accuracy of the software, we manually assigned a set of frames with neutral expression as baseline for the detector. Furthermore, the images / frames identified by this software as including our target AUs were manually FACS coded to confirm the results. Identified images were disregarded if: i) the sequence did not contain the target AUs, ii) the sequence contained AUs with intensities lower than the target AUs, and iii) the sequence was visually difficult to discriminate because of its short duration (<100 ms).

3.5.5 Frequency of Production

We computed the frequency of the target facial expression in the identified segments described above. We defined the frequency of the \( i^{th} \) video fragment as \( f_i = \frac{d_i}{f_r} \), where \( d_i \) is the number of consecutive frames with the target facial expression and \( f_r = 29.97/frames/s \) is the camera’s sample rate.

3.6 Sentences and Questions Used in Experiments 2–4

Tables 3.3–3.4 the seventy-eight (78) sentences and four (4) questions used in Experiment 2. Experiment 3 only includes data for the seventy-eight sentences given in Tables 3.3–3.5 .
Table 3.3: First few sentences used in Experiments 2 & 3. The first column shows the sentence signed / uttered by participants. The second column indicates the polarity of the sentence (positive or negative). The last column indicates the type of sentence (yes/no questions, wh-questions, and assertions).

<table>
<thead>
<tr>
<th>Clause</th>
<th>Polarity</th>
<th>Construction type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brad is cooking fish on the grill</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>Brad is not cooking fish on the grill</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td>Can John eat the last piece of cake?</td>
<td>Positive</td>
<td>Yes/no-question</td>
</tr>
<tr>
<td>Is Mary starting the grill?</td>
<td>Positive</td>
<td>Yes/no-question</td>
</tr>
<tr>
<td>Is Sarah having a party next weekend?</td>
<td>Positive</td>
<td>Yes/no-question</td>
</tr>
<tr>
<td>Is the fork in the kitchen?</td>
<td>Positive</td>
<td>Yes/no-question</td>
</tr>
<tr>
<td>John cannot finish eating the last piece of cake</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td>John is bringing chocolate cake with nuts</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>John is bringing the chocolate cake with nuts</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>John is not bringing the chocolate cake with nuts</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td>Mary is not starting the grill</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td>Mary is starting the grill</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>Sarah had a party yesterday</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>Sarah is going to have a party next weekend</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>Sarah is going to have a party tomorrow</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>The fork is in the kitchen</td>
<td>Positive</td>
<td>Assertion</td>
</tr>
<tr>
<td>The fork is not in the kitchen</td>
<td>Negative</td>
<td>Assertion</td>
</tr>
<tr>
<td>When is Sarah going to have a party?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Where is the fork?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Who can finish the last piece of cake?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Who can’t eat the last piece of cake?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Who is bringing the chocolate cake with nuts?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Who is cooking the fish on the grill?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Who is starting the grill?</td>
<td>Positive</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Why John is not bringing the chocolate cake with nuts?</td>
<td>Negative</td>
<td>Wh-question</td>
</tr>
<tr>
<td>Why Mary has not started the grill?</td>
<td>Negative</td>
<td>Wh-question</td>
</tr>
</tbody>
</table>

Table 3.4: Questions used in Experiment 2.

1. A study shows that tuition should increase 30% for in-state students. What do you think?
2. Some professors think the University should take a stand on the death penalty, gun control and abortion. Do you think that’s a good idea?
3. A group of football fans propose to eliminate exams on Fridays before a home game. Do you agree?
4. Some students think that undergraduates that wish to talk to a professor during office hours should pay a $1 fee. What’s your take on that?
<table>
<thead>
<tr>
<th>Hypothetical conditional clause</th>
<th>Polarity and construction type of each of its parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you are not going to do anything tomorrow, Sara is going to have a party</td>
<td>Positive Hypothetical + Positive Yes/no-question</td>
</tr>
<tr>
<td>If there is still cake left, John can eat it all up</td>
<td>Positive Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If the fork was left in the kitchen, I will get it</td>
<td>Positive Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If Sarah has a party tomorrow, do you want to come to her house?</td>
<td>Negative Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If Sarah has a party tomorrow, I can't go</td>
<td>Negative Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If Sarah has a party tomorrow, I will bring cake</td>
<td>Positive Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If Sarah has a party tomorrow, who is going to bring the food?</td>
<td>Positive Hypothetical + Positive Wh-question</td>
</tr>
<tr>
<td>If the fork is in the kitchen, I won't get it</td>
<td>Positive Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If the fork is in the kitchen, I won't eat it</td>
<td>Positive Hypothetical + Positive Assertion</td>
</tr>
<tr>
<td>If the fork is in the kitchen, I won't get it</td>
<td>Positive Hypothetical + Positive Yes/no-question</td>
</tr>
<tr>
<td>If you are not going to do anything tomorrow, Sara is going to have a party</td>
<td>Positive Hypothetical + Positive Assertion</td>
</tr>
</tbody>
</table>
Chapter 4

Computational Tools for the Analysis of Nonmanuals

4.1 Summary

We developed computational tools that will benefit ASL nonmanuals studies. In particular we developed a deformable shape detector that allows to find salient features, such as the outline of the eyes, eyebrows, mouth and non-salient points such as points in cheeks. The detections found using this algorithm will allows to the trajectories, i.e., the change in facial expressions, over a period of time. To perform inference over this continuous trajectories, we developed a maximum margin classification approach.

4.2 Salient and Non-Salient Fiducial Detection using a Probabilistic Graphical Model

4.2.1 Introduction

Deformable shape detection is an important problem in computer vision and pattern recognition. However, standard detectors are typically limited to locating only a few salient landmarks such as landmarks near edges or areas of high contrast, often conveying insufficient shape information. This chapter presents a novel statistical pattern recognition approach to locate a dense set of salient and non-salient landmarks in images of a deformable object. We explore the fact that several object classes exhibit a homogeneous structure such that
each landmark position provides some information about the position of the other landmarks. In our model, the relationship between all pairs of landmarks is naturally encoded as a probabilistic graph. Dense landmark detections are then obtained with a new sampling algorithm that, given a set of candidate detections, selects the most likely positions as to maximize the probability of the graph. Our experimental results demonstrate accurate, dense landmark detections within and across different databases.

Shape detection is an important problem in computer vision and pattern recognition with applications in recognition, tracking, and classification, amongst others. The goal is to accurately detect the $2D$ position of specific shape landmarks, or *fiducials*, in an image. Some applications such as 3D reconstruction and the recognition of facial expressions require that the deformable shape be described by a *dense set of salient and non-salient* landmarks for satisfactory results [102, 18, 70]. Unfortunately, current detection algorithms are typically tailored to locating only a few salient landmarks [50, 67, 48]. Some exceptions are the 3-Dimensional Morphable Model (3DMM) [20] and the $3D$ model of [66] which find a dense set of face landmarks. However, these methods require a $3D$ database to construct the model and, thus, cannot be learned directly from an image collection.

In this chapter, we propose a novel statistical pattern recognition algorithm to accurately detect a dense set of salient and non-salient landmarks in an image. Our methodology does not require $3D$ object databases and can be used to design landmark detectors for different types of objects – *e.g.*, faces, hands, and structures in medical images. Our approach utilizes the fact that many object classes exhibit a homogeneous structure such that any detected landmark provides contextual information that facilitates the detection of the other landmarks. For instance, Fig.4.1(a) shows a classical example with human faces as objects of interest. While at first sight one may not perceive the 10 faces in this image, once a few fiducials have been detected (*e.g.*, an eye or the nose), the remaining facial parts become readily apparent. Thus, the location of a fiducial automatically provides information on where to find the others.
Figure 4.1: Can you find the 10 faces in (a)? These faces are difficult to see until a face feature is detected (e.g., a nose); then the entire face becomes salient. (b) The output of a standard face landmark detector is typically restricted to a few salient points. (c) Our novel method provides dense detections that include both salient and non-salient landmarks.

In our new method, the relationship between every pair of landmark positions is encoded by the edges of a probabilistic graphical model, where each node represents a landmark position and its local texture. The local texture information of salient landmarks allow them to be detected reliably, whereas non-salient landmark detection is unreliable from just the local texture. Fortunately, the reliable detections constrain the position of non-salient landmarks and vice versa. In addition, the coarsely localized non-salient fiducials aid estimation of other non-salient and misdetected landmarks. As a key result, our detection algorithm can robustly estimate the positions of fiducials in areas such as face cheeks, where a simple local feature detector would generally fail, Fig.4.1(b)-(c). Hence, the resulting algorithm can be used to estimate dense landmark maps in 2D images as in 3DMMs, without requiring prior 3D models.

Our proposed methodology is depicted in Figure 4.2. Using our graphical model (presented in Section 3), landmark detection amounts to maximizing the joint probability of the graph’s nodes in an image. To accomplish this, we propose a sampling algorithm which selects the most likely fiducial positions given a set of candidate detections (Section 4), which resolves the classical computational complexity problem of pattern recognition al-
gorithms that employ graphical models. This algorithm can deal with missing detections, false positives, and occlusions. In addition, we show how to augment the graph and use the original low-level detections to infer many more landmark coordinates in an incremental fashion. Our experimental results demonstrate accurate, dense landmark detections within and across different databases (Section 5).

4.2.2 Related Work

Algorithms such as Active Appearance Models (AAM) [34] and 3DMM [20] use a probabilistic shape and texture model to interpret an object image. These models are learned from a set of annotated training samples so they cannot detect variations beyond what is specified in the training set. In addition, the global shape model often favors a configuration
similar to the mean shape and fails to capture subtle important changes such as eye blinks or single eyebrow motion. Furthermore, these algorithms are best suited to subject-specific modeling. On the other hand, algorithms that rely on fiducial detections are able to fit salient key points reliably without being overconstrained by a global model. However, these approaches can yield unrealistic shape estimates when the global shape is not constrained. These methods have advanced considerably in recent years, with algorithms that even rival human manual annotations [41, 109, 175, 89, 143, 130]. However, they provide a very limited number of fiducials around salient features such as the eyes, nose, and mouth, and most require high-quality images.

Our new pattern recognition method overcomes the above shortcomings by utilizing the positive aspects of fiducial detection and probabilistic shape and texture model approaches. We take advantage of advances on local feature detection to ensure that subtle shape changes are not missed by our method. Each landmark position takes into account the position of all other locally-detected fiducials to generate a plausible configuration for the whole set of detected points. If some landmarks cannot be detected or are misdetected by the local feature detector, the other fiducials will constrain estimation of their positions.

Graphical models have previously been used to guide the fiducial position estimation, although using different approaches which are limited to a very small number of landmarks. Felzenszwalb and Huttenlocher [50] use a tree structured graph to infer the location of 5 face fiducials. For tree based graphs, a poorly localized root node negatively influences all daughter nodes. In contrast, our model represents a dense interconnection of landmarks. Every node has influence from all other shape landmarks so the effect of a few poorly localized nodes is circumvented by information from other nodes in the graph. In another work, Everingham et al. [48] model the joint probability of 9 fiducial positions in faces using a mixture of Gaussian trees. Fiducial are found using a discriminative model with Haar-like features. This algorithm only detects stable points (salient features) and the graph is solely used for robustness to different poses. Gu and Kanade [67] use local feature
detections to generate a set of candidate positions for each fiducial, and then select the most probable set of fiducials using a Bayesian model which encodes the object pose and shape. Their algorithm is initialized using an AAM and a set of fiducials are localized around each landmark. Our algorithm also relies on a set of local feature detections but the highly inter-connected structure of our graph allows us to infer positions of non-salient features while being more robust to occlusions. In general the proposed method outperforms these probabilistic graphical models because it can infer potentially dozens of fiducial positions including non-salient ones. In addition, it easily extends to non-face objects.

4.2.3 Our Probabilistic Graphical Model

Many natural objects are highly structured. Human faces, for example, exhibit strong relationships between the positions of different parts of the face; one can estimate the right eye position reliably given the position of the left eye and the nose by symmetry. In this and other objects, the information from each detected fiducial can be used to infer all other landmarks including those which are unseen, misdetected, or poorly localized. Given this insight, an appropriate modeling scheme describes the pairwise relationship between all fiducials in an object as well as the global configuration. We model the affinity between two fiducials $i$ and $j$ using a potential function $\Phi_{ij}$, and the global configuration using the potential function $\beta$. $\beta$ ensures that the global configuration is reasonable. The logical way to combine these sources of information is using a probabilistic graph.

We define the joint probability of $p$ fiducial positions as,

$$P_{\text{pos}}(X) = \beta(X) \frac{1}{Z_{\text{pos}}} \prod_{i=1}^{p} \prod_{j=1}^{i-1} \Phi_{ij}(x_i, x_j), \quad (4.1)$$

where $X$ encodes the set of coordinates $x_1, ..., x_p$, where $x_i \in \mathbb{R}^2$ has the 2D coordinate of fiducial $i$, $Z_{\text{pos}}$ is called the partition function which makes (4.1) behave as a probability density function (pdf), and $\beta(X) = \exp\left(-\frac{1}{2}(\hat{X} - \mu)^T \Sigma^{-1}(\hat{X} - \mu)\right)$ is a potential function that depends on the Mahalanobis distance from the translated and scale invariant shape $\hat{X}$ to the mean...
shape, \( \mu \). To obtain \( \hat{X} \), we centered the shape to the origin and then normalize by its Frobenius norm. Assuming that our training set has \( N \) shapes \( \hat{X}_1, \ldots, \hat{X}_N \), we estimate the mean of training set \( \mu \) as follows:

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} \hat{X}_i
\]

and the covariance matrix of the training set \( \bar{\Sigma} \) as follows:

\[
\bar{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\hat{X}_i - \mu)(\hat{X}_i - \mu)^T.
\] (4.2)

The parameter \( \alpha \in [0,1] \) controls the penalty of differing from the mean shape. Intuitively, shapes with less deformations prefer a higher value of \( \alpha \), while shapes with larger deformations favor a smaller \( \alpha \).

Assuming that the object is detected and scaled to a standard size, the displacement between two fiducials can be modeled as a bivariate normal distribution. Although displacement between different pairs of fiducials may vary in scale, the correlations may be the same. Therefore, it is important to use a normalized distance, such as the Mahalanobis distance, when measuring the displacement. Thus, the potential functions \( \Phi_{ij}(\cdot, \cdot) \) are defined as

\[
\Phi_{ij}(x_i, x_j) = \exp \left( -(1 - \alpha)\bar{w}_{ij}(\Delta_{ij} - \mu_{ij})^T\Sigma_{ij}^{-1}(\Delta_{ij} - \mu_{ij}) \right),
\] (4.3)

where \( \bar{w}_{ij} \) is the weight of the edge between fiducials \( x_i \) and \( x_j \), \( \Delta_{ij} = (x_i - x_j) \in \mathbb{R}^2 \) is the 2D pairwise distances between landmarks \( i \) and \( j \) for a particular shape, and the parameters \( \mu_{ij} \) and \( \Sigma_{ij} \) are the sample mean and sample covariance of \( \Delta_{ij} \) which are estimated from the training data.

The relationship between landmark positions is encoded as pairwise distances to make the model translation invariant. Since the quadratic term in \( \Phi_{ij}(\cdot, \cdot) \) is a Mahalanobis distance, it is maximum when \( \Delta_{ij} = \mu_{ij} \), and monotonically decreases from there.

In the case of face shapes, the edge connecting the eye to the eyebrow should have a larger weight than the edge connecting the eye to the mouth because the eye more strongly constraints the position of the eyebrow than the mouth. Therefore, we scale the Mahalanobis distances by normalized positive scalar edge weights \( \bar{w}_{ij} \in [0,1] \), to account for this type of variation. The
Normalized edge weights are defined as:

\[ \bar{w}_{ij} = \frac{w_{ij}}{\sum_{k=p}^{l<k} \sum_{k=1}^{l} w_{kl}} , \]

where the edge weights \( w_{ij} \) specify the relative importance of the edges in the graph.

Because the graph encodes the pairwise structure of the shape, an edge connecting nodes \( i \) and \( j \) should have a large weight, \( w_{ij} \), if knowing the position of node \( i \) strongly constrains the position of node \( j \). Also, a large \( w_{ij} \) means the relationship between fiducials \( i \) and \( j \) will be emphasized relative to pairs with smaller weights in (4.3). That is because a larger \( w_{ij} \) forces the optimization algorithm to put more effort into bringing the pairwise distance \( \Delta_{ij} \) closer to its mean value, \( \mu_{ij} \), to maximize the potential. The degree to which \( i \) constrains \( j \) is specified by the sum of the eigenvalues of the covariance matrix \( \Sigma_{ij} \) because the eigenvalues describe the variance along the principal axes of the joint distribution \( N(\mu_{ij}, \Sigma_{ij}) \). A larger variance means we know less about the relative positions of fiducials; the distance between connected fiducials can take on a very wide range of values. However, a smaller variance means the relative position of fiducial \( j \) can be inferred with more certainty. Therefore, we set the weights as:

\[ w_{ij} = \frac{1}{\|\Sigma_{ij}\|_F} . \]

Notice that \( \|\Sigma_{ij}\|_F \) equals the square root of the sum of the eigenvalues of \( \Sigma_{ij} \).

The graph described by equation (4.1) is sufficient to constrain shape estimates to be reasonable, but it is also important to consider the local texture. This is achieved by weighing the pdf by the probability that each landmark is correctly localized, where each texture implies a probability of correct localization. This texture weighting implies that the textures for each fiducial are assumed to be conditionally independent. The graph describing the joint pdf of the landmark positions and textures is defined as,

\[ P(X) = \frac{1}{Z} \beta(X) \prod_{i=p,j<i}^{i=p,j<i} \Phi_{ij}(x_i, x_j) \prod_{k=1}^{P} \gamma_k(x_k), \]  

where \( \gamma_i(\cdot) \) is a potential function of the \( i^{th} \) fiducial’s local texture. More formally, these functions \( \gamma_i \) are the normalized confidences of the local detection. We calculate that confidence using a kernel based density estimation algorithm [63], and normalize such that \( \gamma(\cdot) \) behave as a probability.
4.2.4 Testing Procedure

Fiducial Detection

Fiducial position estimation relies on a set of local detections. The local detections are a result of evaluating a classifier for each fiducial at each image patch within the image regions expected to contain the associated fiducial. The regions are determined by the training data. The classifiers rely on features extracted from the image patches, and are based on the local texture or context features. Salient points like corners of the eyes prefer local texture (pixel value features) as in [41], while non-salient points such as points in the cheeks require a more global texture feature [137]. The classifiers are learned using Kernel Linear Discriminant Analysis (KLDA) [108, 15] and the annotated training data. KLDA is a pattern recognition technique that finds a low-dimensional projection of the image features $x \in \mathbb{R}^d$ which maximally separates samples from different classes while grouping members of the same class. We use the Radial Basis Function (RBF) kernel defined as:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

where $\sigma$ is a parameter to select, and $x_i$ are the vectorized sample images. The kernel parameter is optimized using the criterion of [172].

Each local detector returns a set of candidate positions $D_i = \{\hat{x}^1_i, ..., \hat{x}^{m_i}_i\}$, where $\hat{x}^j_i$ is the $j^{th}$ candidate position of fiducial $i$, and $m_i$ is the number of candidate detections for fiducial $i$. These candidate detections consisting of their position and corresponding texture provide information about the unknown true fiducial positions in the graph. The main task of this model is to find the set of candidate detections that maximize the graph probability, given their position and local texture. More formally, the objective is

$$X^* = \arg\max_{x_1, ..., x_p} P(X), \; s.t. \; x_i \in D_i, \; i = 1 ... p.$$  \hspace{1cm} (4.5)

Estimation of the Fiducials

Estimating fiducial positions amounts to optimizing the objective function of equation (4.5). The size of our search space depends on the number of candidates per fiducial. This space is very large and evaluating the probability of each configuration requires $\prod_{i=1}^p m_i$ trials. In the ideal case, we would
like to evaluate the probability of all possible combinations. This is computationally intractable, so we generate a reduced set of probable configurations from the full set of candidates. Instead of uniformly testing candidate configurations, we generate configurations that are highly probable given our probabilistic model. This allows us to find a reasonable shape without testing every possible case. The most likely configuration initializes the second stage, a maximum likelihood propagation algorithm that fine tunes the estimate to maximize the proposed model.

To find a set of likely fiducial configurations, the optimization algorithm assumes that at each iteration, \( p - 1 \) “known” fiducials (salient and non salient) are coarsely detected. We then randomly sample for the \( p^{th} \) fiducial knowing that its pdf is conditioned by the “known” ones. This process is iterated for all fiducials until a likely global configuration is found. More formally, to find the \( \hat{x}_i, i = 1 \ldots p \), which maximizes (4.4), we initialize \( x_i \) randomly to one of the local detections in \( D_i, i = 1 \ldots p \). In the first stage, we sequentially update \( x_k, k = 1 \ldots p \), by taking a random sample from the conditional pdf of \( x_k \) given all other nodes:

\[
P(x_k|\cdot_{\neq k}) \propto \gamma_k(x_k) \exp \left( -\frac{1}{n} \sum_{j \neq k} \bar{w}_{kj}(\Delta_{kj} - \mu_{kj})^T \Sigma_{kj}^{-1}(\Delta_{kj} - \mu_{kj}) + \frac{\alpha}{2} x_k^T \bar{\Sigma}_{kk}^{-1} \hat{x}_k \right),
\]

where \( \cdot_{\neq k} = \{x_1, x_2, \ldots, x_{k-1}, x_{k+1}\ldots x_p\} \), \( \hat{x} \) is the \( k^{th} \) fiducial of \( \hat{X} \), \( \Delta_{kj} = x_k - x_j \) and \( \bar{\Sigma}_{ij}^{-1} \) is the inverse of the submatrix that measures the covariance of the \( i^{th} \) and \( j^{th} \) fiducial of scale and translation invariant shape \( \hat{X} \) in Eq. 4.2. This procedure is repeated for all the fiducials \( x_k \) to generate a sample \( \hat{X}_i \), where \( i = 1, \ldots, M \), where \( M \) is the maximum number of iterations. Note that the conditional pdf of equation (4.6) can be calculated explicitly and normalized for the set of discrete candidate positions. We store the final configuration and the corresponding probability from equation (4.4) after every iteration.

To be robust to empty and false positive candidates resulting from occluded or noisy images, we use the probabilistic graph of equation (4.1) to augment the sets \( D_i \) with the maximum-likelihood (ML) estimate of each fiducial position \( x_i \) given \( x_{-i} \). Taking the logarithm of equation (4.1) yields...
a quadratic function. Taking derivatives, we find that

\[ x_k^{MLE} = ((1 - \alpha)G_p + \alpha G_s)^{-1}((1 - \alpha)W_p + \alpha W_s) \quad (4.7) \]

where

\[
G_p = \left( \sum_{j \neq k} \tilde{w}_{kj} \Sigma_{kj}^{-1} \right)^{-1},
\]

\[
G_s = s_c \Sigma_{kk}^{-1}, \quad s_c = \|X\|^2,
\]

\[
W_p = \left( \sum_{j \neq k} \left[ \tilde{w}_{kj} \Sigma_{kj}^{-1} (x_j - \mu_{kj}) \right] \right),
\]

\[
W_s = s_c \left( \Sigma_{kk}^{-1} \mu_k + \sum_{j \neq k} \Sigma_{kj}^{-1} (\hat{x}_k - \mu_k) \right).
\]

\( \mu_k \) is the \( k^{th} \) coordinate of \( \mu + t_x \), where \( t_x \) is the centroid of \( X \). Note that this procedure without taking into account \( x_k^{MLE} \) is commonly known as a Gibbs Sampling [57].

In the second phase of the optimization, we initialize with the previously sampled configuration \( \tilde{X}_i \) that maximizes equation (4.4), and repeat the sequential procedure until convergence or a maximum number of iterations. However, instead of drawing random samples from the conditional pdf, we select the value \( x_k \in D_k \) which maximizes equation (4.4). The procedure is summarized in algorithm 1.

### 4.2.5 Experiments

We performed a series of experiments on face, heart, and hand-shape detection to demonstrate the utility of the proposed approach. The parameter \( \alpha \) which controls the amount of regularization was cross validated separately for the different object types. The parameter \( M \) determines how many exemplars are obtained in the first stage of inference. We found experimentally that \( M = 100 \) was sufficient to initialize the maximum likelihood propagation algorithm near a local minimum of (4.4).
Algorithm 1 Inference Algorithm

Input=\{D_1, ..., D_p\}
for \(i = 1\) to \(p\) do
  Set \(x_0^i\) = random sample of \(\tilde{D}_i\)
end for
for \(stage = \) sampling to maximum likelihood propagation do
  for \(t = 1\) to \(M\) do
    for \(i = 1\) to \(p\) do
      Set \(\tilde{D}_i = \{D_i, x_0^i, \text{MLE}_i\}\)
      Calculate \(P(x_{t-1}^i|x_{t-1}^{t-1})\) using (4.6) and the candidates \(\tilde{D}_i\)
      if \(stage = \) sampling then
        Let \(x_t^i\) be a random sample of \(P(x_t|x_{t-1})\)
      else
        Let \(x_t^i\) be \(\arg \max_{\tilde{x}_i^t \in \tilde{D}_i} P(x_{t-1}^i|x_{t-1}^{t-1})\)
      end if
      Let \(x_{t-1}^i \rightarrow x_t^i\)
    end for
    if \(x^i\) did not change then
      end maximum likelihood propagation
    end if
  end for
Set \(x^0\) as the \(x^i\) that maximize (4.4)
end for

Face Landmark Detection

We first evaluate the proposed algorithm on three face databases: AR [101], Labeled Faces in the Wild (LFW) [78], and the XM2VTS database [105]. Faces are roughly localized in position and scale using the Viola-Jones face detector, then cropped and scaled to 150 × 150 pixels, corresponding to a mean inter-eye distance of approximately 15 pixels. For the AR database, we train with 448 images, and test on 448 images containing subjects not found in the training set. For the LFW database, we train with 1027 images, and test on 500 images of subjects not used in training. For the XM2VTS database, we train with 448 images, and test on 350 images containing subjects not found in the training set. We also use the model trained with the LFW database to detect fiducials on the AR and XM2VTS databases. The error is measured as the Euclidean distance from the ground truth fiducial positions (i.e., manual markings) to the estimated position for a total of 50 fiducials over all test images. Results are compared to those of the AAM algorithm.

Example detections using our algorithm are shown in Fig. 4.3 with error histograms shown in Fig. 4.4. The first three histograms are for within database testing. That is, the training and testing set, although disjoint, are from the same database. The last three histograms are for across database experiments. In this case, we employed the training set of one database and do the testing on the images of the other specified database. These are the most challenging and interesting experiments.
Figure 4.3: We show example results of the derived approach. From top to bottom, the rows correspond to the shape detections for the AR database [101], the LFW database [78], and the XM2VTS database [105]. The database contains face images in unconstrained environments. Results show robustness to occlusions, pose, and lighting.

Figure 4.4: Error histograms (Euclidean distance, in pixels) for a total of 50 detected fiducial points versus the ground truth of the testing sets. The ground truth positions were obtained by manual annotation. PGA denotes our new probabilistic graph algorithm, while AAM denotes the classical AAM.
Figure 4.5: Comparison of our algorithm with AAM [34] and the local detector of [48]. Our method provides more precise detections than the AAM, and many more fiducial points than the local feature detector.

Figure 4.5 shows an additional visual comparison between our method, [34] and [48]. Our algorithm outperforms the AAM in every case. For the case of constrained databases such as the XM2VTS and AR databases, our algorithm slightly outperforms the AAM. However, for the case of the unconstrained LFW database, and the across-database experiments, our algorithm performs significantly better than the AAM. This reiterates the problem of learning a global probabilistic shape and texture model. It also verifies the benefit of using a probabilistic graphical combined with local detection of fiducials.

We also compare our method to the algorithm of [48] on the AR and LFW databases for the 9 fiducial points detected by their algorithm. The implementation was obtained from the authors’ website, and was already trained. On the AR database, our algorithm achieves an error with mean and standard deviation of $1.5315 \pm 1.2751$ pixels while [48] achieves an error rate of $1.9189 \pm 1.4619$ pixels. On the LFW database, our algorithm achieves an error with mean and standard deviation of $2.5052 \pm 1.6972$ pixels, while [48] obtains an error rate of $2.2801 \pm 4.1774$ pixels. On the XM2VTS database, our algorithm achieves an error with mean and standard deviation of $1.5561 \pm 1.7515$ pixels, while [48] obtains an error rate of $1.7922 \pm 1.8497$ pixels. In summary our method provides more accurate detections and additional set of non-salient landmarks.
Hand-Shapes

To further challenge the algorithm, we detected 52 landmarks describing the hand contour. The image database of [139] contains 40 images of 4 subjects showing different hand-shapes. We rescaled the images to 100 × 100 pixels, then for 3 folds of cross validation, we randomly selected 30 images for training and 10 images for testing. Fig. 4.6(a) shows examples of the detected landmarks and the approximate contour of the hand.

Following the same comparison procedure of Section 4.2.5 and combining the 30 testing results of all cross validation folds, our algorithm outperforms AAM with a mean error and standard deviation of 1.8805 ± 2.3707 pixels while AAM achieves an error rate of 3.5008 ± 3.0955 pixels. The error histograms are shown in Fig. 4.6(b).

Incremental Learning: Inferring Additional Landmarks

So far we have shown two cases where the algorithm can be used for detecting shapes having salient and non-salient landmarks. This is sufficient for most applications, but some require a shape representation with many more landmarks. Given the assumption that a set of localized fiducials constrains the position of other shape landmarks, we should be able to accurately predict the position of $m >> p$ fiducials based on a localized subset of $p$ fiducials obtained using the above approach. To achieve this, note that equation (4.7) only relies on the shape part of the graphical model (no reliance
on texture). Therefore, we can learn the parameters for an $m$ node graph from an annotated set of images, and simply estimate the position of each of the undetected fiducials as the MLE estimate of that fiducial position given the position of the $p$ previously detected fiducials using equation (4.7). Fig. 4.7 shows the much denser shape detection achieved using the 50 fiducial detections of Fig. 4.3 and an augmented shape model. Note that no local feature detection or iterative sampling was done to infer the much denser set of landmarks.

4.2.6 Conclusion

We have presented a new statistical pattern recognition algorithm for detecting a dense set of salient and non-salient landmarks in images of a deformable object. This method exploits the fact that each landmark position constrains the position of other landmarks on an object. In our model, the pairwise relationships between landmarks is naturally encoded in the nodes and edges of a probabilistic graph. Given a set of candidate landmarks provided by local detectors, our algorithm selects the set of candidate detections that maximize the joint probability of the graph. Besides being more accurate, our method also provides dense detections of deformable shapes and is not restricted to faces. The face experiments show detection of significant local deformations of the eyes.
and mouth, while the hand experiments demonstrate precise detection with global deformation. Furthermore, results including training and testing across different datasets show that our method outperforms specialized, salient feature detectors as well as AAMs.

4.3 Maximum-Margin Functional Classification in Computer Vision

4.3.1 Introduction

Although many computer vision problems are defined by a functional feature space, most classification methods derived to date assume sample feature vectors are characterized in a vector space. For example, in the recognition of gate, basketball actions or facial expressions, it is common to define each sample as a set of image features (e.g., SIFT, SURF, magnitude of the optic flow) and then derive advance statistical classification algorithms to learn to discriminate between samples representing distinct classes. A more natural feature space to describe such activities is one where objects are characterized as functions defining the movement of a set of fiducial points moving in space and time (i.e., their natural movement observed in a video sequence). The present chapter introduces a statistical classification method, based on the maximum-margin approach, to maximize the separation of continuous functions of different classes without the need of any prior knowledge of the functional space or the problem. Specifically, we propose to identify the derivatives of the sample and basis functions maximizing the separation between classes. The derived algorithm generalizes previous versions of functional Support Vector Machines (SVM) and has several advantageous properties. Experimental results with challenging datasets demonstrate the derived algorithm readily adapts to different problems, including, sparsity, occlusions, and small number of features.

A natural way to characterize actions, activities and object categories seen in video sequences is to employ functions defining the space-time movement of the object’s fiducial points [147, 146, 169]. For example, in the recognition of facial expressions, one can use a function to define the spatio-temporal movement of each face landmark (e.g., the left-most corner of the left eyebrow), Fig. 4.8. Yet, to date, the most used approach to define sample videos is in the form of a vector space, e.g., the response of a set of Gabor filters on the facial landmark in Fig. 4.8 [97, 49], or the procrustes shape of the deformation of the object [68, 71]. This is so, not because this vector representation is more
Figure 4.8: Shown here are the two functions $x_i(t)$, $i = 1, 2$, defining the spatio-temporal movement of two fiducial points in a video sequence of an American Sign Language sentence. The feature vector of the sample video is the vector of functions $F = [x_1(t), x_2(t)]^T$. In this model, our goal is to find the linear differential equations $L(F(t))$ resulting in the representation with largest possible margin between samples of different classes.

desirable than its functional counterpart, but because the most advanced maximum-margin-based classification algorithms are derived for such static feature representations, not for functional ones. This chapter aims to remedy this shortcoming.

A common way to model these functional representations is by using discriminative or generative learning machines with feature vectors defining samples at different time intervals of the functions (typically called bins) [76], i.e., a discretization of the temporal data. This sampling approach has serious disadvantages. Consider, for example, the movement of the landmarks in videos of facial expressions, Fig. 4.8. In this application, this discrete sampling approach would require equal number of frames across sample videos [76]. In order to ensure this same length limitation, a naive solution is to either truncate or neglect frames or align videos using, iterative approaches [176] or learning-based methods [156]. Occlusions pose a major challenge to these solutions. Landmarks generally become occluded at some point of the video sequence, adding errors in the alignment process. This makes an approximate modeling of the object even less accurate, which eventually translates in lower classification accuracies [94].

The key idea of the present work is to identify a norm on the original functional space $H$ which
intrinsically maps the original representation in $\mathcal{H}$ to another space $\mathcal{F}$ where the functions can be readily separated using the maximum-margin criterion. In particular, we use the inner product of the derivatives of the basis functions of $\mathcal{H}$ to non-linearly map sample functions of different classes into a space where these are linearly separable. One advantage is that the dual space can be characterized by the distribution of the derivatives of degree smaller or equal than $v$ [25]. Note that the inner product of the derivatives of these basis functions correspond to local deformations of $\mathcal{H}$, allowing the algorithm to linearize the classification boundary, i.e., it can be used to define the norm that maximizes the margin. We derive the primal and dual solutions for this problem, define its properties, and show experimental results on new and standard datasets.

### 4.3.2 Basic Formulation

Functional Data Analysis (FDA) [128, 51, 170] is the collection of techniques employed to model objects best defined as continuous functions. FDA generally uses the filtering approach [128], which models observations as a linear combination of a set of basis functions with a specified level of smoothness. An appropriate norm is defined to derive the approach. This is particularly useful to derive classification algorithms.

In [128] a functional linear discriminant analysis is defined by finding the within and between scatter matrices in the functional space given by a set of basis functions defined a priori. A similar method is proposed in [128] to adapt optimal scoring [73]. Penalized Discriminant Analysis [72] is also used to find discriminant functions with the desired level of smoothness, preventing rank deficient matrices and overfitting. On the nonparametric side, [51] approximates the observations using Parzen function estimation, rather than relying on the use of basis functions. Support Vector Machines (SVM) [145] for functional data have also been derived. In [149], the authors propose a linear and a Kernelized functional SVM by using the coefficients of the filtering approach. In [100], a $L_1$-norm maximum margin is derived to find a sparse separating function.

All the aforementioned methods can be applied to either the observations or their derivatives or linear combinations of them. However, these algorithms are specially tailored to deal with problems consisting of a single observation. Most problems in computer vision relay on many continuous observations, e.g., multiple landmarks in dynamical facial expressions as shown in Fig. 4.8. This limits the use of these techniques to a small number of applications. Also, the above listed algorithms...
need to be tailored to each problem, \textit{i.e.}, it is imperative to know \textit{a priori} which linear combination of derivatives leads to a reasonable functional representation of the data. This is made clear by the small number of applications where the above mentioned algorithms have been successfully utilized.

The present work derives a functional classification algorithm that maximizes the separation of continuous functions of different classes \textit{without} prior knowledge about the norm or differential equation that best separate samples of different classes. Our proposal is to simultaneously find the coefficients of a differential equation that maximize the separation between classes and the curve that best separates the functions. To achieve this, we employ the maximum-margin criterion of SVM. The resulting method is highly interpretable – it is easy to identify the components of the functions that carry the most discriminant information. Furthermore, the proposed method is not constrained to classification problems described by single continuous functions, but can deal with multiple functions defined in distinct functional spaces. The trick is to identify the derivatives that are most appropriate for each sample function. Thus, our approach does not need to make use of intermediate representations [177] to represent complex actions, making it more robust to model complex actions.

To study our solution in detail, let us start with a formal definition of the problem. A functional SVM attempts to find the function \( w(t) \) that maximizes the separation between sample functions belonging to two classes [145]. Let us define the training samples of a functional binary classification problem as the set \( \{(\gamma_i(t), y_1), \ldots, (\gamma_n(t), y_n)\} \), where \( \gamma_i(t) \in \mathcal{H}^v \), \( \mathcal{H}^v \) a Hilbert space of continuous functions with bounded derivatives up to order \( v \), and \( y_i \in \{-1, 1\} \) their class labels. When the samples of distinct classes are linearly separable, the function \( w(t) \) that maximizes class separability is given by

\[
J(w(t)) = \min_{w(t)} \frac{1}{2} \langle w(t), w(t) \rangle_{\mathcal{F}},
\]

s.t. \( y_i(\langle w(t), \gamma_i(t) \rangle_{\mathcal{F}} - b) \geq 1, \quad (4.8) \)

where \( b \) is the bias and \( \langle \gamma_i(t), \gamma_j(t) \rangle_{\mathcal{F}} = \int \gamma_i(t) \gamma_j(t) dt \) denotes the functional inner product.

Typically, a few outliers prevent us from finding the separating hyperplane. This problem is easily addressed by inclusion of the slack variables \( \xi = [\xi_1, \ldots, \xi_n]^T [145] \), yielding the following optimization approach

\[
90
\]
\[ J(w(t), \xi) = \min_{w(t), \xi} \left\{ \frac{1}{2} \langle w(t), w(t) \rangle_F + C \sum_{i=1}^{n} \xi_i \right\} \]
\[ \text{s.t. } y_i(\langle w(t), \gamma_i(t) \rangle_F - b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad (4.9) \]

where \( C > 0 \) is a penalty value found using cross-validation.

### 4.3.3 Functional Differential Vector Machines

We are interested in learning the linear differential operator that maximizes classification. Let us define this linear differential operator \( L(.) \) as

\[ L(\gamma(t)) = \psi(t) = a_0 \gamma^{(0)}(t) + a_1 \gamma^{(1)}(t) + \ldots + a_v \gamma^{(v)}(t) \]
\[ = \sum_{k=0}^{v} a_k \gamma^{(k)}(t), \quad (4.10) \]

where \( \gamma^{(i)}(t) = \frac{d^i \gamma}{dt^i} \) is the \( i^{th} \) derivative of the function \( \gamma(t) \). Our goal is to jointly find the set of coefficients \( a = [a_0, \ldots, a_v]^T \) and the function \( w(t) \) that maximally separate the training samples of different classes.

The linear differential inner product is obtained by using (4.10) to define \( \langle \cdot, \cdot \rangle_{L(F)} \),

\[ \langle \gamma_i(t), \gamma_j(t) \rangle_{L(F)} = \sum_{k,l=0}^{v} a_k a_l \int \gamma_i^{(k)}(t) \gamma_j^{(l)}(t) dt \]
\[ = \sum_{k,l=0}^{v} a_k a_l \langle \gamma_i^{(k)}(t), \gamma_j^{(l)} \rangle_F. \quad (4.11) \]

Substituting (4.11) in (4.9) yields
\begin{equation}
J(\omega(t), \xi, \mathbf{a}) = \min_{\omega(t), \xi, \mathbf{a}} \left\{ \frac{1}{2} \sum_{k,l=0}^{v} a_k a_l \langle \omega^{(k)}(t), \omega^{(l)}(t) \rangle_F + C \sum_{i=1}^{n} \xi_i + \lambda \| \mathbf{a} \|^2_2 \right\},
\end{equation}

\text{s.t. } y_i \left( \sum_{k,l=0}^{v} a_k a_l \langle \omega^{(k)}(t), \gamma_i^{(l)}(t) \rangle_F - b \right) \geq 1 - \xi_i
\xi_i \geq 0,
\end{equation}

where \( w(t) = \sum_{k=0}^{m} a_k \omega^{(k)}(t) \) and \( \lambda \) penalizes the 2-norm of the coefficients \( \mathbf{a} \).

### 4.3.4 Functional Differential Vector Machines with Basis Functions

FDA usually assumes the functions are constrained to spaces spanned by a set of \( m \) basis functions \( \phi(t) = \{ \phi_0(t), ..., \phi_{m-1}(t) \} \), with \( \gamma_i(t) = \sum_{r=0}^{m} c_i^r \phi_r(t) \), where \( \mathbf{c}_i = [c_0^i, ..., c_{m-1}^i]^T \) is the vector of coefficients. This is sometimes called the filtering approach [128]. Using the linear combination approach, we can redefine the inner product by replacing it in (4.11), yielding

\begin{equation}
\langle \psi_i(t), \psi_j(t) \rangle_{L(F)} = \sum_{k,l=0}^{v} a_k a_l \int \phi_i^{(k)}(t) \phi_j^{(l)}(t) dt
= \sum_{k,l=0}^{v} a_k a_l \Phi_{k,l} \mathbf{c}_i^T \mathbf{c}_j \end{equation}

where \( \Phi_{k,l}^{(k,l)} \) is a \( m \times m \) matrix with elements \( \Phi_{k,l}^{(k,l)} = \langle \phi_i^{(k)}, \phi_j^{(l)} \rangle_F \). Letting \( \omega(t) = \mathbf{w}^T \phi(t) \), and applying Lagrange to (4.12) yields the primal optimization problem,

\begin{equation}
\mathcal{L}(\mathbf{w}, b, \xi, \mathbf{a}, \alpha) = \min_{\mathbf{w}, \xi, b, \mathbf{a}, \alpha} \left\{ \frac{1}{2} \mathbf{w}^T \left( \sum_{k,l=0}^{v} a_k a_l \Phi(k,l) \right) \mathbf{w} + C \sum_{i=1}^{n} \xi_i + \lambda \| \mathbf{a} \|^2_2 \ight. \\
- \sum_{p=1}^{n} \alpha_p \left( y_p \left( \mathbf{w}^T \sum_{k,l=0}^{v} a_k a_l \Phi(k,l) \mathbf{c}_p - b \right) - 1 + \xi_p \right) - \sum_{p=1}^{n} \theta_p \xi_p \right\}.
\end{equation}
where $\alpha_p$ and $\theta_p$ are Lagrange multipliers.

### 4.3.5 Dual Optimization Problem

The dual optimization problem is found by computing the partial derivatives with respect to the variables in (4.14),

\[
\frac{\partial L}{\partial b_p} = \sum_{p=1}^{n} \alpha_p y_p = 0, \\
\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \theta_i = 0, \\
\frac{\partial L}{\partial w} = \left( \sum_{k,l=0}^{v} a_k a_l \Phi^{(k)(l)} \right) w - \sum_{k,l=0}^{v} a_k a_l \Phi^{(k)(l)} \sum_{p=1}^{n} \alpha_p y_p c_p = 0.
\]

From (4.17) we have $w = \sum_{p=1}^{n} \alpha_p y_p c_p$, i.e., the representer theorem [155]. Taking partial derivatives with respect to the coefficients $a$ gives us,

\[
\frac{\partial L}{\partial a_i} = a_i w^T \Phi^{(i)(i)} w \\
+ \frac{1}{2} \sum_{l=0, l \neq i}^{v} a_l w^T \left( \Phi^{(i)(l)} + \Phi^{(l)(i)} \right) w + 2 \lambda a_i \\
- \sum_{p=1}^{n} \alpha_p y_p \sum_{l=0, l \neq i}^{v} a_l w^T \left( \Phi^{(i)(l)} + \Phi^{(l)(i)} \right) c_p \\
- 2 a_i \sum_{p=1}^{n} \alpha_p y_p w^T \Phi^{(i)(i)} c_p = 0.
\]

Solving for $a$ in (4.18) and using the representer theorem we obtain $Ma = \lambda a$, where $M$ is a symmetric $v \times v$ matrix given by

\[
M_{ij} = \begin{cases} 
\frac{1}{2} w^T \Phi^{(i)(i)} w, & \forall i = j \\
\frac{1}{2} w^T \left( \Phi^{(i)(j)} + \Phi^{(j)(i)} \right) w, & \forall i \neq j.
\end{cases}
\]

Note that $a$ cannot be disassociated from the primal solution, whereas $w$ is expressed as a linear combination of the training vector $c_p$. Thus, we can jointly estimate $\alpha$ and $a$ using the
wrapper approach \cite{127, 80}, \textit{i.e.}, iteratively minimizing over $a$ and maximizing over $\alpha$. Substituting (4.15),(4.16) and (4.17) in (4.12) yields the solution

$$D(\alpha, a) = \max_\alpha \min_a \sum_{p=1}^n \alpha_p - \frac{1}{2} \sum_{p,d=1}^n \alpha_p \alpha_d y_p y_d \langle c_p, c_d \rangle \Phi + \lambda \|a\|_2^2, \quad (4.19)$$

where

$$\langle c_i, c_j \rangle = c_i^T \left( \sum_{k,l=0}^v a_k a_l \Phi(k,l) \right) c_j. \quad (4.20)$$

\subsection*{4.3.6 Multidimensional Functional Differential Vector Machine}

As stated earlier, most computer vision applications require the simultaneous modeling of multiple functions. This is readily achieved in our formulation by using the definitions in (4.10) and the approach derived in Section 4.3.4.

Let us define the multidimensional sample function as $\gamma(t) = [\gamma_1(t), \ldots, \gamma_g(t)]^T$, and the linear differential operator as $L(\gamma(t)) = [L_1(\gamma_1(t)), \ldots, L_g(\gamma_g(t))]^T$, where $L_i(\cdot)$ is defined as in (4.10). \textit{Importantly, each function $\gamma_i(t)$ can have a different underlining hypothesis, \textit{e.g.}, $\gamma_1(t)$ can be described with Fourier descriptors, and $\gamma_2(t)$ can be described as a compact support by spline bases.}

Let $\bar{c} = [(c^1)^T, \ldots, (c^g)^T]$, where $c^i$ is the vector of linear coefficients for the $i^{th}$ function. The multidimensional inner product is defined as

$$\langle \gamma_i(t), \gamma_j(t) \rangle_L = \sum_{e=1}^g \langle c^e_i, c^e_j \rangle_{\Phi_e}, \quad (4.21)$$

where $\langle c_i, c_j \rangle_{\Phi_e}$ is the inner product defined in (4.20) with respect to the $e^{th}$ basis. Note that this new definition will need a set of coefficients $\bar{a} = [(a^1)^T, \ldots, (a^g)^T]^T$, where $a^i$ are the linear coefficients of $L_i(\cdot)$. Following the same procedure as that given in Section 4.3.4, we find the dual optimization problem,

$$D_M(\alpha, \bar{a}) = \max_\alpha \min_{a^i} \sum_{p=1}^n \alpha_p - \frac{1}{2} \sum_{e=1}^g \sum_{p,d=1}^n \alpha_p \alpha_d y_p y_d \langle c^e_p, c^e_d \rangle_{\Phi_e} + \sum_{e=1}^g \lambda_e \|a^e\|_2^2. \quad (4.21)$$
Notice that, as defined in (4.20), the minimization order of the $a_i$’s is irrelevant.

### 4.3.7 Finite Differences

In some cases, the derivatives of the functions need to be estimated, because there is either no hypothesis on the form of the functions or these are not differentiable, e.g., wavelets. Fortunately, numerically estimated derivatives are a direct result of the derivations in Section 4.3.4.

Let $\mathbf{x} = [x(t_0), x(t_1), \ldots, x(t_m)]^T$ be a vector whose elements are the samples of a continuous function at times $t_0, t_1, \ldots, t_m$ and let $\Delta t = |t_j - t_{j-1}|$, $j = 2, \ldots, m$. The estimation of the discrete derivative using Euler’s finite differences method is

\[
\frac{dx}{dt} \approx D\mathbf{x}, \quad (4.22)
\]

\[
D = \frac{1}{\Delta t} \begin{bmatrix}
-1 & 1 & 0 & \ldots & 0 & 0 \\
0 & -1 & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & -1 & 1 \\
0 & 0 & 0 & \ldots & 0 & -1
\end{bmatrix}_{m \times m}
\]

The dual optimizations defined in (4.19) and (4.22) can be done by approximating the matrix $\Phi^{(k)(l)} \approx (D^k)^T D^l$.

### 4.3.8 Experiments

In this section we show the benefits of the proposed Maximum Margin Functional Classification approach with synthetic and real data.

#### Synthetic Data

In this experiment, we compare the proposed approach with linear SVM [145]. We synthetically generated samples of two classes, where each class is described by three functions. Specifically,
samples drawn from class 1 have the following form:

\[
x_{1k} = \begin{bmatrix}
\cos(2\pi t_1 + .2) + \epsilon_{11}^k \\
\sqrt{t_2} + \epsilon_{12} \\
\exp \left((t_3 - 5)^2\right) + \epsilon_{13}
\end{bmatrix}, \quad t_1 = [0 1 10 \ldots 4], t_2 = [0 1/40 \ldots 1] \text{ and } t_3 = [0 1/4 \ldots 10].
\]

While samples from class 2 are of the following form:

\[
x_{2k} = \begin{bmatrix}
\cos(2\pi t_1) + \epsilon_{11} \\
\sqrt{t_2} + \epsilon_{12} \\
\exp \left((t_3 - 5.1)^2\right) + \epsilon_{13}
\end{bmatrix}, \quad t_1 = [0 1/10 \ldots 4], t_2 = [0 1/40 \ldots 1] \text{ and } t_3 = [0 1/4 \ldots 10],
\]

where \( x_{ik} \in \mathbb{R}^{3\times40} \) is the \( k^{th} \) sample of the \( i^{th} \) class (\( i = 1, 2 \)) and \( \epsilon_{lj}^k \in \mathbb{R}^{40} \), \( l = 1, 2, 3 \), \( j = 1, 2 \), is a zero-mean random Gaussian vector. Note that both classes have similar functions except that the first and third functional dimensions are slightly shifted between classes 1 and 2.

**Robustness to Missing Data**

First, we test the claim that a functional classification approach is more robust to missing data because the missing information is intrinsically given by the estimation of the underlying function(s).

To this end, we randomly select a percentage of samples and assign them missing values to a single random column of \( x_{ik} \). We generated 100 samples for each class with a variance of \( .2 \). For the functional SVM, we use the basis function approach described in Section 4.3.6, where the first, second and third functional dimensions are given by the Fourier, polynomial and \( \beta \)-spline bases, respectively, and we used derivatives up to order 2. For linear SVM, we vectorized the samples so that \( \hat{x}_{ik} = \text{vec}(x_{ik}) \in \mathbb{R}^{120} \) and we eliminate the features that had missing elements. We varied the percentage of samples with missing entries, and for a given percentage we performed 20 iterations of 10-fold cross-validation. As shown in Fig. 4.9, the classification accuracy drops very faster for linear SVM. In comparison, the recognition accuracy of the derived approach remains nearly unaffected.
Figure 4.9: Classification accuracy of synthetic data for different percentage of samples with missing data. As the number of samples with missing data increases, the classification accuracy for the proposed approach remains nearly unchanged. In comparison, the classification accuracy of a linear SVM drops quite fast.

Robustness to Noise

Next, we measure how the classification accuracy is affected by the noise variance $\epsilon_{i,j}^k$. It is important to emphasize that neither the derived approach nor traditional SVMs are robust to noise. Our goal is not to define a robustified version of the derived algorithm. Rather, the goal is to show that a reasonable estimate of the true, unknown underlying function(s) will be more robust to noise than traditional classifiers.

We generated 100 samples for each class. For the proposed method, we used the same approach defined in Section 4.3.8. For linear SVM, we vectorized the samples so that $\hat{x}_{i,k} = vec(x_{i,k}) \in \mathbb{R}^{120}$. We changed the variance $\epsilon_{i,j}^k$ between 0 and 1, and for a given noise we performed 20 iterations of 10-fold cross-validation. As seen in Fig. 4.10 the classification accuracy drops significantly faster for linear SVM than for the derived algorithm.

Real Data

We present results on the novel and very challenging dataset of [16], the recent database of [54], and the standard set of [171].
Nonmanuals in ASL

American Sign Language (ASL) utilizes two components, namely the manuals (i.e., handshape, hand movement and palm orientation [111]) and the nonmanuals (i.e., body posture, head orientation and facial expressions). We use the database of ASL nonmanuals of [16], Fig. 4.11(a). This dataset contains 344 videos of 9 ASL native deaf subjects signing hypothetical conditionals (e.g., If I were a fish) in each of the two polarities, i.e., positive and negative.

The pre-processing of these videos included: face detection using the algorithm of [150], detection of facial landmarks, i.e., contours of the eyes, eyebrows, nose, mouth and jaw line, using the non-linear active appearance model of [69], and the decoupling of the motion and shape of the face using the non-rigid structure from motion algorithm of [62]. We selected 19 two-dimensional landmark points describing the shape of the eyes, eyebrows and mouth (as described above) plus the rotations of the head, i.e., yaw, pitch and roll. This yields a 41 multidimensional functional space.

The following reasons make this database extremely challenging: i) **Sparsity.** Most of the landmark’s spatio-temporal information does not contribute to the classification of hypotheticals. Rather, most of the movement of landmarks are specific to the subject. Our goal is to demonstrate that the proposed algorithm can identify a sparse metric able to focus on the few discriminant functional features in this data. ii) **Occlusions.** Several of the tracked face landmarks become occluded at some point in the video. The derived algorithm has to adapt the metric to avoid using outliers or
Facial Expressions of Emotions

We study the categorization of dynamic facial expressions of emotions. In particular, we used five of the prototypical expressions of emotion [44], i.e., the facial expressions of happiness, surprise, sadness, anger and disgust. We test the three-dimensional dynamic facial expression database of [171]. This dataset consists of 606 video sequences. Each emotion category is expressed by 101 human subjects. As in the section above, we used 19 two-dimensional landmarks describing the shape of the eyes, eyebrows and mouth, for a total of 38 multidimensional functional feature space.

This database does not contain occlusions. What makes this dataset challenging is the use of such a small number of facial landmarks for the categorization of emotions. To achieve good classification accuracies in this database, the algorithm needs to use the information of all the functions defining the movement of the face landmarks. This is in fact the opposite of the sparsity case studied above.
We used a one-versus-one approach [56], where each class-pair is compared with the other. A 10-fold cross-validation test yields a classification accuracy of 77%. This is in comparison with the use of the classical functional SVM which did not converged and thus did not yield any meaningful result.

**Gesture Recognition**

We study the categorization of twelve gestures captured by a 3D Microsoft Kinect camera as described in [54]. This database consists of 594 samples of 12 different gestures, Fig. 4.11(c). The categories are: lift with outstretched arms, duck, push right, put goggles, wind it up, shoot, bow, through, had enough, change weapon, beat both and kick. Each frame is defined by twenty three-dimensional points describing the motion of the gesture, so that a sample is represented by a 60-dimensional functional vector. The functions describing each coordinate of the joints were regressed finding the beginning automatically by measuring if the energy with respect to the previous frame was above a learned threshold. The challenge of this database is in the larger number of classes used.

We performed a leave-four-samples-out cross-validation test, i.e., using all the samples but 4 for training, and the rest for testing. As above, we used the one-vs-one classification approach [56], training $12(12-1)/2$ classifiers and using voting to make the final decision. The average classification rate was 96.8%. This result is much better than the 76.5% accuracy reported in [54] using a Random Forest classifier.

**4.3.9 Conclusions**

We have derived the primal and dual solutions for the maximum-marging classification of functions defining the spatio-temporal movement of image landmarks. The key contribution of the chapter is to make use of the derivatives of the basis functions (of the functional space describing the sample features) to non-linearly map the original representation into a space where the original functions are linearly separable. The trick is to only define this mapping intrinsically by means of an inner-product defining the norm of the original functional space. This is equivalent to the kernel-trick used in machine learning. The main difference is in the use of the derivatives of the functions, which allows us to change the norm locally for each of the sample functions. This trick allows the use of our
approach in complex computer vision problems by simultaneously modeling the multiple functions that define the action or object under study.

The present work detailed the theoretical arguments and derivations of the proposed approach and showed how the derived algorithm can be used to simultaneously deal with multiple functions. Although this is mostly a theoretical section, we have also presented experimental results in three classical computer vision problems. These experimental results demonstrated the usefulness of the derived approach even when the number of tracked fiducial points is small, the dataset contains significant occlusions, or the discriminant information is sparse. These results thus demonstrate the versatility of the derived approach.
Chapter 5

Conclusions and Future Work

5.1 Conclusions

We showed a methodology to uncover the discriminant features and temporal structures of the linguistic model governing nonmanuals in sign languages. As it was shown in Chapter 2 a linguistic representation of the face is obtained, followed by a computational approach to determine the combination of features consistently observed in each class but not with others. The proposed methodology proved to be able to discriminate between nine different classes of sentences – Hypothetical conditionals, Wh-questions, Wh-questions postposed and Assertions in their two polarities and Yes/no questions in positive polarity.

Our results suggested differentiator factors in the frequencies used in head-shakes for assertions and hypothetical conditionals. In addition, the results also highlight the role that the mouth and teeth play in negation, conditionals and Wh-questions in ASL, something that was just hypothesized before our work.

We identified a facial expression of negation used consistently across cultures as a non-verbal gesture and in languages as a co-articulator and a grammatical marker. Our results suggest that the “not face” might have an origin in facial expressions of emotion. This also suggests an evolutionary path for the development of non-verbal communication and grammatical markers. We found the same production in non-verbal gesturing (without speech) in people of distinct cultures, as a co-articulator in speech (i.e., English, Mandarin and Spanish), and, across modalities as a grammatical marker in ASL. We showed that in the case of ASL a facial expression can be grammaticalized, something that was just researched in the case of headshakes [122].
We presented an algorithm for detecting a dense set of salient and non-salient fiducials in faces and the methodology can be easily extended to images of other deformable objects such as hands and medical images. This was done by taking advantage of the fact that many landmarks are highly constrained by the positions of other landmarks, something that fits with the definition of a undirected graphical model. In particular, given a set of candidate landmarks provided by local detectors, our algorithm selects the set of candidate detections that maximize the joint probability of the graph that defines global shape and local texture. The face experiments show accurate detection of local deformations of parts of the face such as eyes and mouth, while experiments such as the hand demonstrated precise detection of globally deformable objects.

Additionally, we develop a new algorithm to classify continuous functions, an algorithm that would benefit detection of sign language facial events. In particular, we have derived the primal and dual solutions for the maximum-margin classification of functions defining the spatio-temporal movement of image landmarks. Instead of using the functions themselves, we also made use of the derivatives of basis functions to non-linearly map the original representation into a space where the original functions are linearly separable. The work detailed in Chapter 4 gives the theoretical arguments and derivations to simultaneously deal with multiple functions. Although this is mostly a theoretical section, we have also presented experimental results in three classical computer vision problems.

5.2 Future Work

The purpose of this study is to understand how head and facial nonmanuals define the grammar in American Sign Language. To do so, we first hypothesized that temporal relations encode well studied nonmanuals as well as unknown nonmanuals. Unfortunately this method requires the data to be already labeled frame by frame, a task that is labor intensive. To fully extend our work for the analysis of complex constructions and continuous signing, we need a tool that can provide these annotations automatically. To automatically annotate the linguistic-computational model, we need to develop new tools to confront different challenges. Empirical evidence has shown that with the current technology, some linguistic components (e.g., blinks) cannot be characterized merely by using the shape. We investigated this phenomenon by using state-of-the-art algorithms. Methods such as AAM based approaches [68, 168] are relatively robust to occlusions but the precision decays
drastically. Other probabilistic graphical model methods [17] can improve the precision but they are not robust to large occlusion and the computational complexity is very high. Even though the aforementioned methods suffer from major disadvantages, all of them can roughly detect areas of interest such as mouth, nose, eyes and eyebrows. We can then combine the appearance, i.e., the pixels surrounding the area of interest, and the shape to determine their current state. Another problem is data imbalance, i.e., the number of samples per category is disproportionally different, and the dimensionality of the data can be very high and the number of samples can be very large (so called Big Data). In order to alleviate the aforementioned problems, we will need to use additive models [74] by subsampling the training set to have the same number of samples for each one of the linguistic model categories. Also, the model should include temporal information by modeling the interaction between different labels by using generative or discriminative models (HMM or CRF).

Additionally, the development of a novel functional multidimensional maximum margin approach can be applied to different domains besides ASL nonmanuals, e.g., facial expression of emotions, shape analysis, action recognition, electroencephalogram (EEG) signals, among many others. Furthermore, as noted before, the solution for the functional SVM can be interpreted as relaxed formulation of Multiple Kernel Learning [127], in order to complete the theoretical framework of maximum margin approaches for functional data, we need to provide further evidence of generalization, bounds for risk minimization and a multiclass Elastic Net type SVM approach as in [88].
References


105


