Image and Vision Computing 32 (2014) 671–681

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions☆

Jingjing Liu [a,1], Bo Liu [a,1], Shaoting Zhang [b,*], Fei Yang [a], Peng Yang [a], Dimitris N. Metaxas [a], Carol Neidle [c]

[a] *Department of Computer Science, Rutgers University, Piscataway, NJ, USA*
[b] *Department of Computer Science, University of North Carolina at Charlotte, NC, USA*
[c] *Linguistics Program, Department of Romance Studies, Boston University, Boston, MA, USA*

ARTICLE INFO

Article history:
Received 7 June 2013
Received in revised form 26 December 2013
Accepted 21 February 2014
Available online 12 March 2014

Keywords:
American Sign Language (ASL)
Non-manual grammatical markers
Eyebrow height
Head gestures
Facial expressions
Conditional Random Field (CRF)

ABSTRACT

Changes in eyebrow configuration, in conjunction with other facial expressions and head gestures, are used to signal essential grammatical information in signed languages. This paper proposes an automatic recognition system for non-manual grammatical markers in American Sign Language (ASL) based on a multi-scale, spatio-temporal analysis of head pose and facial expressions. The analysis takes account of gestural components of these markers, such as raised or lowered eyebrows and different types of periodic head movements. To advance the state of the art in non-manual grammatical marker recognition, we propose a novel multi-scale learning approach that exploits spatio-temporally *low-level* and *high-level* facial features. *Low-level* features are based on information about facial geometry and appearance, as well as head pose, and are obtained through accurate 3D deformable model-based face tracking. *High-level* features are based on the identification of gestural events, of varying duration, that constitute the components of linguistic non-manual markers. Specifically, we recognize events such as raised and lowered eyebrows, head nods, and head shakes. We also partition these events into temporal phases. We separate the anticipatory transitional movement (the *onset*) from the linguistically significant portion of the event, and we further separate the *core* of the event from the transitional movement that occurs as the articulators return to the neutral position towards the end of the event (the *offset*). This partitioning is essential for the temporally accurate localization of the grammatical markers, which could not be achieved at this level of precision with previous computer vision methods. In addition, we analyze and use the motion patterns of these non-manual events. Those patterns, together with the information about the type of event and its temporal phases, are defined as the high-level features. Using this multi-scale, spatio-temporal combination of low- and high-level features, we employ learning methods for accurate recognition of non-manual grammatical markers in ASL sentences.

1. Introduction

Signed languages are full-fledged natural languages, comparable in structure and complexity to spoken languages but manifested in the visual–gestural modality. Computer-based recognition of sign language from video, which has been the object of various research efforts over the last two decades (e.g., [42,49,47]), is particularly challenging in that it requires attention to detection and interpretation of linguistic information conveyed through both the manual and the non-manual channels. Although the equivalents of words are articulated primarily through movements of the hands and arms, important linguistic information of various kinds is also expressed non-manually: through head movements and facial expressions. In particular, essential grammatical information about such things as negation, clausal type, question status, and topics is conveyed by clusters of specific facial gestures and head movements [3–5,9,22,23,34]. For instance, the marking of yes/no questions typically includes raised eyebrows. However, raised eyebrows are a component of many other grammatical markers, as well (e.g., topics, conditional clauses, "relative clauses"); these expressions are distinguished by other differences in facial expressions, including eye gaze and eye aperture, nose configuration, and head positions and movements.

Despite their linguistic importance, it is only relatively recently that sign language recognition has begun to focus on detection of the non-manual components of signed languages, in some cases as an aid to the recognition by computer of manual signs [2,39], and in other cases, with a view toward interpretation of the grammatical information carried by non-manual markers [26,30,37]. Computer-based recognition of non-manual aspects of signed languages has made use of methods for facial expression recognition [44,13] and head pose estimation [32], which have been active research topics in computer vision for

☆ This paper has been recommended for acceptance by Qiang Ji.
* Corresponding author. Tel.: +1 732 4450636; fax: +1 732 4450537.
E-mail address: szhang16@cs.uncc.edu (S. Zhang).
[1] Indicates equal contribution.

http://dx.doi.org/10.1016/j.imavis.2014.02.009
0262-8856/© 2014 Elsevier B.V. All rights reserved.

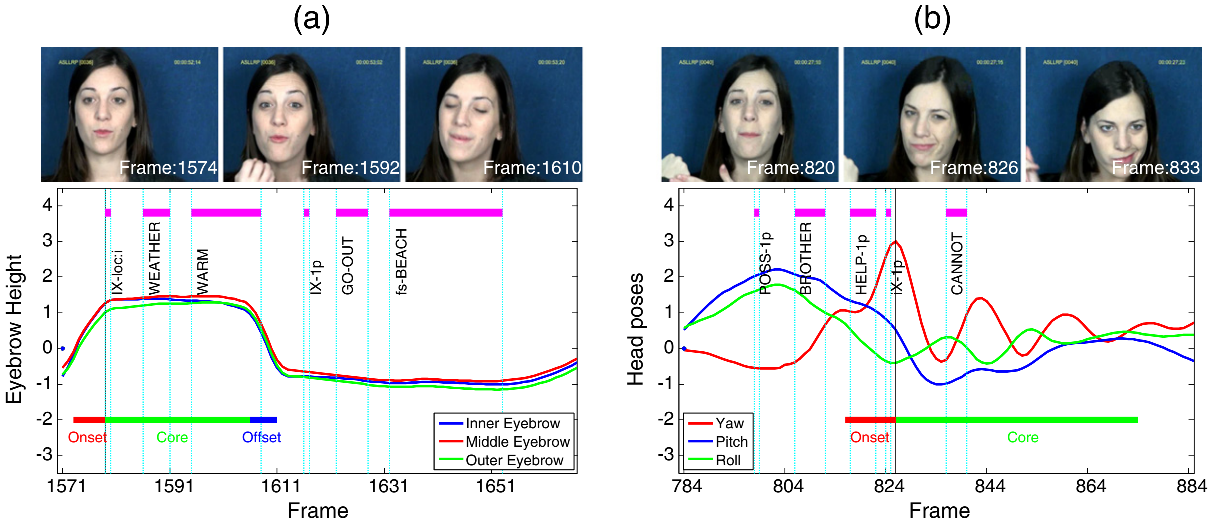

Fig. 1. Eyebrow raise and head shake: (a) If the weather is warm, I'll go to the beach; (b) As for my brother helping me, he can't.

decades. Most approaches to interpretation of non-manual information have involved construction of mapping functions between low-level features and grammatical markers. Related work will be reviewed in Section 2. However, the accurate interpretation of non-manual events (e.g., eyebrow gestures and periodic head movements) that carry linguistic information requires attention to the spatio-temporal patterning of these events over syntactic domains that may vary significantly in size and duration.

In this paper we propose an automatic recognition system for non-manual grammatical markers based on multi-scale and spatio-temporal analysis of head poses and facial expressions. Although there has been previous research focused on low-level features alone [26,30,36], or on high-level features that can be derived from those [37], our approach relies on combining both types of features. To extract the relevant features, we use a 3D deformable face tracker based on an adaptive ensemble of ASMs [17]. This 3D tracker deals well with the large head movements and occlusions of the face by the hands that occur during signing. Another advantage of the 3D face model approach is that it eliminates the need for facial pose alignment necessary in 2D approaches, which is often a source of significant errors in facial pose estimation and expression recognition. Another feature of our 3D face tracker is that it is generic and not person-specific. Our face tracker can adaptively fit each person's face without the need for person-specific training. The low-level features related to the 3D head pose and facial expressions (including eye aperture and eyebrow configuration) are based on geometry and appearance features from the 3D face tracker. The low-level features are used, in part, to recognize non-manual events such as raised/lowered eyebrows and periodic head movements (nods and shakes). The recognition of these events allows us to extract important high-level features, including the type of non-manual event, its temporal phases, and specific aspects of its motion patterns over phrasal domains. Finally, we combine the low- and high-level features based on learning methods [1] to recognize the five types of non-manual grammatical markers under consideration in this paper.

We pay particular attention to the computer-based recognition of raised and lowered eyebrows, and of periodic head nods and shakes, since they function as components of many non-manual grammatical markers in signed languages generally, and in American Sign Language (ASL) in particular. The focus on eyebrow gestures and periodic head movements is thus linguistically motivated. We further partition these events into temporal phases, to facilitate the accurate localization of the relevant grammatical markers. For example, in eyebrow events, our research to date has revealed that the linguistically significant portion of a raised or lowered eyebrow gesture begins after an anticipatory, transitional phase (which we refer to as the *onset*), during which the eyebrows raise (or lower) to the maximal extent. The *core* of the gesture begins at that point, linguistically aligned with the start of the relevant sign or group of signs associated with the grammatical marker. The eyebrows often start returning to neutral (a phase that we refer to as the *offset*) a few frames before the end of the final sign in the phrase being marked by the relevant non-manual expression. This is illustrated in Fig. 1(a). For periodic head gestures, the head tends to begin with a transitional, anticipatory movement (*onset*) that involves rotating the head to the maximal angle, so that the first head nod or shake can cover the maximal angular range, and the start of the linguistically significant portion of these periodic head movements is usually initiated from this maximally rotated position. Periodic head movements tend to involve successive head rotations of diminishing amplitude, eventually damping out (with no identifiable *offset*). This pattern is illustrated in Fig. 1(b). Hence, it is critical to separate out the preparatory phase in order to focus on the linguistically meaningful part.

In order to recognize these high-level features, we introduce a hierarchical Conditional Random Field (CRF) [20] framework to recognize raised/lowered eyebrows, head nods, and head shakes from video sequences, and to further partition these non-manual events into temporal phases, i.e., onset, core and offset (if there is one). After recognizing the non-manual events, we analyze their motion patterns, which, together with the temporal phase partitioning, are regarded as high-level features. Experiments are designed to evaluate the recognition results for eyebrow gestures and periodic head movements, as well as the further effects of such identifications on successful recognition of non-manual grammatical markers. More specifically, we investigate the improvement in performance for recognition of these markers achieved as a result of the combined use of low-level and high-level features (i.e., eyebrow and head events and their spatio-temporal patterning). Our method outperforms the baseline approach that uses only low-level features. The methodology in this paper is a significant extension of our previous approach in [25]. The innovations here include: 1) utilization of drastically improved facial tracking that combines a 3D deformable model with an ASM approach; this new method is capable of reporting reliably when tracking is not successful; 2) improvement in the computation of high-level features as sequence models, which in turn enhances the incorporation of linguistic knowledge into the recognition of grammatical markers; and 3) validation of the results through a substantial number of additional experiments, demonstrating success in the identification and temporal localization of events and their phases, and in the higher-level detection and identification of grammatical markers in ASL.

This paper is organized as follows. Section 2 reviews related work on the recognition of non-manual expressions and head movements, as well as the recognition of non-manual grammatical markers. Section 3 introduces our proposed method which is based on: 3D face tracking to obtain landmarks; extraction of the relevant low-level features; selection of the relevant facial features with a ranking model; recognition of non-manual events using hierarchical CRFs; and derivation of additional high-level features from motion analysis of these events. We also describe the integration of multi-scale features for non-manual grammatical marker recognition. Section 4 presents the experiments on recognition of eyebrow gestures and periodic head movements, and then on recognition of grammatical markers in ASL. Section 5 summarizes our approach.

2. Related work

In signed languages, essential grammatical information is expressed by various combinations of facial expressions involving the eyes, eyebrows, nose, and mouth, as well as head movements, including periodic head nods and shakes [3–5,9,22,23,34]. For automatic analysis of facial expressions, numerous vision approaches have been proposed. Prior approaches can be categorized into two main methodological types: classification-based and action unit-based. The classification-based methods aim at recognizing a small set of prototypical facial expressions related to certain mental activities [41,7,48,6], as well as identifying the intensities of these expressions [24,38]. Most of these approaches are based on data collected in laboratory environments, with frontal-view faces and few occlusions; these constraints do not hold for sign language data gathered in more natural settings. Furthermore, the targeted expressions are commonly limited to the six universal emotions (i.e., fear, anger, sadness, happiness, surprise, and disgust) proposed by Ekman et al. [11], whereas automatic analysis of facial expressions for sign language recognition must encompass a wide range of expressions involved in conveying linguistic information. The action unit-based approaches analyze facial expressions from the perspective of facial actions related to muscle activity, and describe these expressions by their locations and intensities. In this stream of approaches, the Facial Action Coding System (FACS), which deconstructs facial expressions into specific Action Units (AUs) [12], is widely used. Many previous methods focus on the recognition of these AUs and their temporal phases [43,45]. Although some grammatical facial expressions are related to AUs, little prior research attempts to explore the linguistically significant motion patterns of these actions, nor to use such information for non-manual grammatical marker recognition in ASL.

Other issues relevant to the non-manual channel include head pose estimation and head movement recognition. Out-of-plane head motion is frequent during signing. Hence, it is essential to estimate head pose, and to eliminate the perspective effects that distort facial features. Some models can be used to estimate head poses along with facial landmark fitting, such as the Active Appearance Models (AAM) [8], Active Shape Models (ASM) [17], and Constrained Local Model (CLM) [10]. See [32] for a survey of head pose estimation methods. Some efforts have also been made to recognize head motions, including periodic movements [31,18,14] of the kind that occur frequently as components of non-manual markers. However, sub-classification of these gestures based on their movement properties and differentiation of their temporal phases has not been previously undertaken. In this paper, we show that such analyses provide significant advantages for non-manual grammatical marker recognition.

Recently, researchers have tackled recognition of non-manual grammatical markers through analysis of linguistically relevant facial expressions and head movements [46,35,27]. Some methods have constructed frame-based or sequence-based mapping functions between non-manual grammatical markers and low-level features. Michael et al. [28,29] used SIFT + Pyramids and a bag of words approach to capture facial features. Metaxas et al. [26,30] additionally introduced head pose and some geometric facial features, such as eyebrow height and eye aperture. However, the low-level features utilized in these methods do not explicitly contain important temporal patterns of linguistic significance. Nguyen et al. [36,37] recognized head motions for non-manual grammatical marker recognition, for instance, head forward, move down, or turn left. Although the information about these head motions is in some sense high-level, the authors ignored some linguistically important temporal properties of these head movements. They also did not exploit low-level facial features for recognition of non-manual expressions. In this paper, we focus on the multi-scale, spatio-temporal analysis of facial expressions and head poses, incorporating the temporal properties of non-manual events, and we further leverage this information to enhance non-manual grammatical marker recognition.

3. Our approach

3.1. Overview

The flowchart of our approach for non-manual grammatical marker recognition is shown in Fig. 2. This method involves the following steps:

1. We use a 3D deformable face tracker based on an adaptive ensemble of ASMs [17], which is capable of tracking in the presence of large

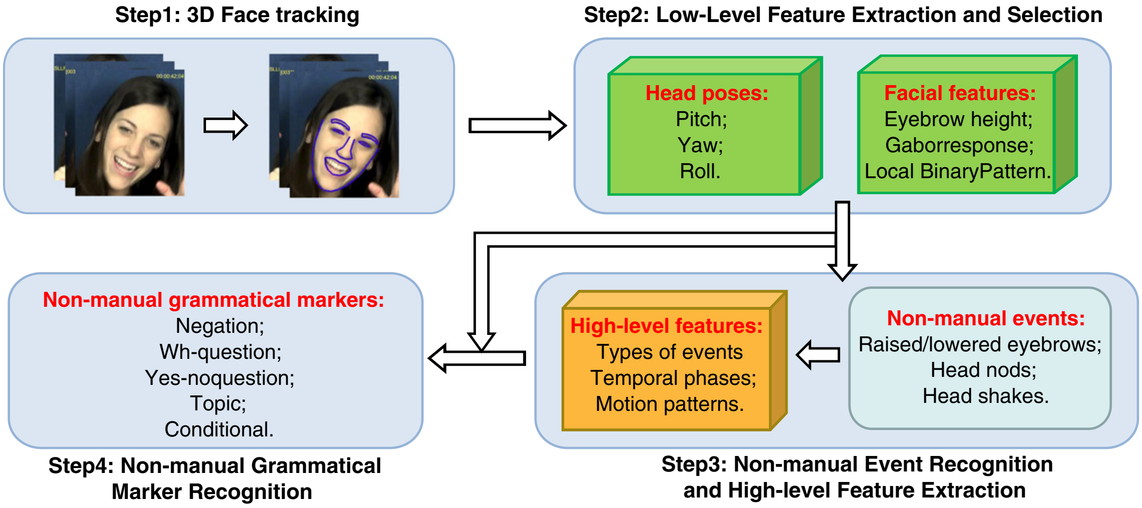

Fig. 2. Flowchart of the proposed method.

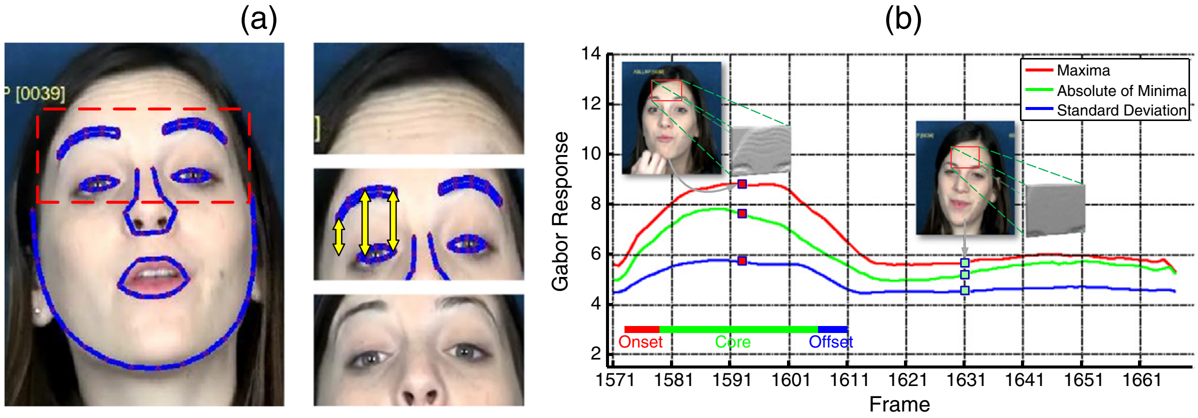

Fig. 3. Low-level features for eyebrow gesture recognition. (a) Geometric and Appearance features. On the right, from top to bottom are forehead region; illustration of inner, middle, and outer eyebrow height shown by yellow lines; ROI for eyebrow gestures. (b) An example of the Gabor response of forehead during a raised eyebrow event.

head movements and occlusions of the face by the hands that may occur during signing. The use of a 3D facial tracking approach eliminates the use of alignment methods necessary in 2D approaches, which often result in inaccurate pose and expression feature estimation. Our parameterized deformable model is then used to extract the features listed below.

2. We extract low-level features related to the 3D head pose and facial expressions, based on geometry and appearance.
3. We first use the low-level features to recognize non-manual events such as raised/lowered eyebrows, head nods, and head shakes. The recognition of these events then allows us to extract important high-level features, such as the type of non-manual event, its temporal phases, and motion patterns over phrasal domains.
4. We combine the low- and high-level features based on learning method to recognize the five types of non-manual grammatical markers under consideration in this paper.

In the following we give details for each of the steps of our approach.

3.2. Step 1: face tracking

Accurate facial tracking is essential for precise non-manual marker feature extraction and detection. We adopt the method in [17], which has been successfully used in many applications (e.g., [26,25]), and combine the multiple ASMs with a 3D deformable model. The face tracker is capable of tracking the non-linear geometry of the facial shape manifold in the presence of large head movements because it uses an ensemble of Point Distribution Model (PDM) clusters as well as a 3D shape model. The combined ASM and 3D face tracker automatically locates 79 facial landmarks, as shown in Fig. 3(a), and estimates the three head rotation angles (i.e., pitch, yaw, and roll). In sign language videos, accurate face tracking is challenging because of occlusions caused by arbitrary head pose and hands. To alleviate this problem, we modify the generic setting in [17]. We enhance the constraint power of the global 3D shape prior while maximizing mixture density probability of local components, for a tradeoff between occlusions and local deformation.

Furthermore, as the model tracks a face, it creates a stored 3D model of the face being tracked. When the face is occluded by the hands, our face model erroneously uses features that include the hands. This results in a distortion of the tracked face parameters, which, as a result, then do not match the stored facial parameters; in such cases, we know that our model is not tracking well at this point. When the occlusion ends, the model again fits well with the facial data, and the stored model's parameters again match those of the tracked model. The frames where our face model does not track well (as a consequence of some type of occlusion or even noise in the data) are not used here for the recognition of non-manual markers. This results in a significant improvement in our recognition results, as compared with our previous work [25].

3.3. Step 2: low-level facial feature extraction and selection

There are two types of low-level features. The first type is extracted from the face tracker based on the rigid 3D head movement; such features include yaw, pitch, and roll, as well as head translation. The second type is based on the geometry and the appearance of the relevant facial expressions, such as eye aperture and eyebrow movement and configuration. For recognition robustness, we extract and combine both geometric and appearance-based facial features.

As shown in Fig. 3(a), the intuitively obvious feature to describe eyebrow gestures is eyebrow height. Based on facial landmarks, the inner, middle, and outer heights of the right and left eyebrows are computed. We also extract the eyebrow Region of Interest (ROI), from which we compute the LBP (Local Binary Pattern) features. In addition, we use texture features from the forehead region. As illustrated in Fig. 3(b), with eyebrow raise, the forehead appears more wrinkled; in contrast, lowered eyebrows are typically accompanied by brow furrowing, resulting in a smooth forehead. We obtain the Gabor features in the forehead regions using vertical filters; we calculate the maxima, minima, and standard deviation of the Gabor response in the ROI. These features have not been used previously for recognition of eyebrow configuration.

The raised and lowered eyebrow movements undergo ordinal changes during the onset and offset phases. Inspired by previous work in facial expression recognition [38] and age estimation [21], which reveal that features or feature combinations carrying rank information can improve the performance of vision tasks involving ordinal changes, we further explore the underlying relationships between LBP features in ROI and eyebrow movements. Intuitively, the feature dimensions whose values monotonically increase (decrease) during onset (offset) can be more discriminative. In this work, we use the Ranking SVM [15] algorithm for LBP feature selection. It aims at finding a hyperplane under pairwise ranking constraints, by maximizing SVM soft margins:

$$
\begin{aligned}
&arg\ \min_{\boldsymbol{w},\xi,}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum\xi_{i,j,k}\\
&\text{s.t.}\forall\left(\boldsymbol{x_i},\boldsymbol{x_j}\right)\in\boldsymbol{r}_1^*:\boldsymbol{w}^T\varphi(\boldsymbol{x_i})\geq\boldsymbol{w}^T\varphi\left(\boldsymbol{x_j}\right)+1-\xi_{i,j,1}\\
&\qquad\cdots\\
&\quad\forall\left(\boldsymbol{x_i},\boldsymbol{x_j}\right)\in\boldsymbol{r}_n^*:\boldsymbol{w}^T\varphi(\boldsymbol{x_i})\geq\boldsymbol{w}^T\varphi\left(\boldsymbol{x_j}\right)+1-\xi_{i,j,n}
\end{aligned}
\tag{1}
$$

where C is a trade-off parameter between training error and soft margin; $\xi_{i,j,k}$ are slack variables; $\boldsymbol{r}_k{}^*(k=1,\ldots,n)$ are targeted ranking lists. $(\boldsymbol{x_i},\boldsymbol{x_j})\in\boldsymbol{r}^*$ means the rank of $\boldsymbol{x_i}$ is higher than $\boldsymbol{x_j}$'s in $\boldsymbol{r}^*$.

More specifically, we estimate the onset and offset phases of eyebrow events from the original feature sequences. Each phase corresponds to one ranking list. A rank value is assigned to each frame in the list according to the frames' temporal order in the eyebrow movement. A larger rank value is assigned to the frame with higher eyebrow height. Take the onset of eyebrow raise as an example. We assign rank 1 to the first frame of onset phase, rank 2 to the second frame, etc. The dataset for LBP feature selection is divided into two groups: D_1 and D_2. D_1 is used to train the ranking model for each individual feature dimension. For M-dimension features, M ranking models are learned using SVM$^{\text{rank}}$ [16]. D_2 is used to evaluate the extent to which features preserve ordinal information. We calculate the ranking score of each feature dimension s_i, $i = 1,\ldots, M$ on D_2. The score is defined as the similarity between the prediction results and the ground truth [19]:

$$\tau\left(\hat{L}, L\right) = \frac{\sum_{i,j=1}^{N}\left(\widehat{L_i}-\widehat{L_j}\right)\left(L_i-L_j\right)}{N(N-1)} \tag{2}$$

where $\lceil\cdot\rceil$ is 1 if the inner function is positive, and 0 if it is negative. $\hat{L}$ and L are the predicted rank list and ground truth, respectively. A higher score indicates greater ability of a feature to describe the eyebrow configurations. In our recognition task, we rank s_i, $i = 1,\ldots, M$ and the Kth largest corresponding features are selected.

3.4. Step 3: non-manual event recognition and high-level feature extraction

Most attempts at grammatical marker classification or recognition have been based on direct construction of mapping functions between low-level features and non-manual grammatical markers. However, low-level features are ill suited to capturing important properties of phenomena that occur over varying spatio-temporal scales, and they are sensitive to small errors in tracking and to noise in the image. In this section, we introduce hierarchical CRFs to recognize important non-manual events—i.e., raised or lowered eyebrows, head shakes, and head nods—and to further partition them into onset, core, and offset phases. Through temporal analysis of these events, we further extract additional temporal properties and patterns in the form of high-level features. These high-level features represent non-manual events explicitly, which contributes significantly to reliable recognition of grammatical markers.

3.4.1. Non-manual event recognition and phase partitioning

To obtain accurate detection of eyebrow or periodic head gestures as well as their temporal phases (i.e., including onsets and offsets, where relevant, as described earlier), sequence models could be utilized to label the frames in a video sequence. Models such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs) have demonstrated advantages for sequence classification and event detection, such as in the sign language domain [37,42] citeASL:yang09. We adopt the CRF approach and construct a two-stage model for non-manual event recognition and partition.

CRF is a probabilistic model proposed by Lafferty et al. [20]; it has been widely used for structured prediction, such as image segmentation, event detection, and object tracking. The model considers not only the dependencies between observations and states, but also interactions among states. As discriminative models, CRFs have several advantages over HMMs. They allow arbitrary dependencies between observations, and they need only a small training dataset, because there is no requirement to specify the distributions of the states and observations.

In a chain CRF model, given an observation sequence X, the probability of a label sequence Y has the form:

$$p(Y|X) \propto \exp\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_i f\left(y_t, x_t^i\right) + \sum_{t=1}^{T}\sum_{j=1}^{M}\mu_j g\left(y_t, y_{t-1}^j\right)\right) \tag{3}$$

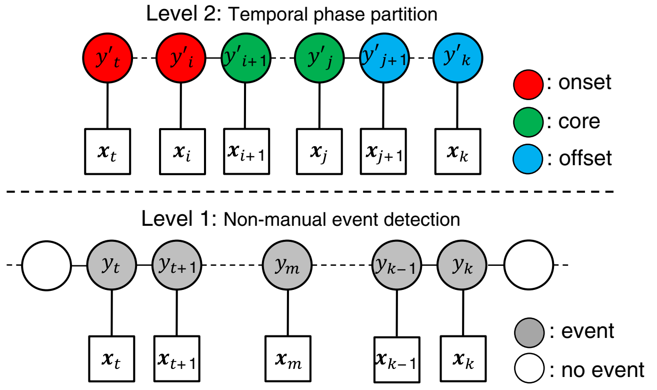

Fig. 4. The 2-level CRFs for recognition of non-manual events and their temporal phases. Each square (x) represents the low-level features of every frame, and each circle corresponds to the frame state. y denotes the first-level state, showing whether a frame is involved in the non-manual event. y' denotes the second-level state, indicating frames belonging to onset (red), core (green), or offset (blue) phases.

where T, N, and M are the numbers of the nodes, feature values, and states, respectively; $f(y_t, x_t^i)$ is the unary potential function to evaluate the interactions between features and labels; and $g(y_t, x_{t-1}^j)$ is the binary potential function considering the dependencies among neighborhood labels. λ_i and μ_j are the parameters we can learn from the training data using likelihood maximization.

We propose a hierarchical CRF framework to detect and partition the non-manual events. This framework is composed of two-level chain CRFs, as illustrated in Fig. 4. At the first level, using low-level features, the CRF model is trained to recognize the entirety of the non-manual event (specific type of eyebrow gesture or head movement). In Fig. 4, one non-manual event is detected in frames t to k whose first-level states are marked as gray circles. Then at the second level, another CRF model is trained to further partition the detected event into temporal phases, using the same low-level features.

For the eyebrow motion CRF models, the unary potential function $f(y_t, x_t^i)$ that we use is feature quantization: the feature value is quantized according to predefined thresholds. For each feature dimension, rather than using evenly divided intervals, we develop an adaptive strategy to decide the thresholds used for quantization, as illustrated in Fig. 5. We first compute the value ranges (blue bars) of lowered, normal, and raised eyebrows in the training set, which establish four of the thresholds: maximum feature values of lowered and normal eyebrow heights, and minimum feature values of normal and raised eyebrow heights (refer to the red dots in Fig. 5). We chose four additional thresholds (refer to green dots), defined as follows: the 80th percentile of lowered eyebrows, the 20th and 80th percentiles of normal eyebrows, and the 20th percentile of raised eyebrows. These eight thresholds are used to discretize the feature values into nine integers.

3.4.2. High-level feature extraction

The high-level features aim to model the motion patterns of each of the distinct non-manual markers, for which the general patterns hold across different ASL signers. In this paper we report our first attempt

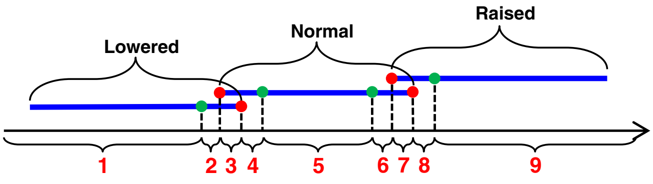

Fig. 5. Statistical feature discretization by leveraging true labels.

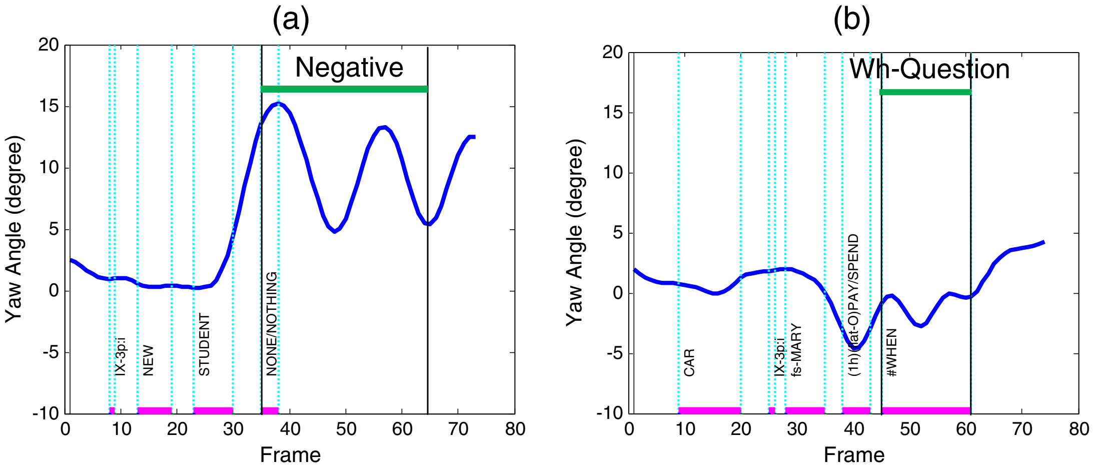

Fig. 6. Example of yaw angle curve in Negation and Wh-questions.

to explore such spatio-temporal patterns in non-manual marker recognition. Detailed motion analysis is conducted for head shake and head nod based on the event detection and the phase partitioning results. As shown in Fig. 6, a given kind of head motion varies significantly as far as its patterning (e.g., with respect to amplitude and velocity) is concerned in different non-manual grammatical markers in ASL. For example, as has been well described in the linguistic literature, the non-manual marking of negation typically includes a head shake with a relatively large amplitude, whereas the head shake that is sometimes found in wh-questions has a much smaller amplitude, but with more rapid repetitions.[2] In order to model this difference, we develop discriminative features as follows.

First, we detect the "peak frame" points from the overall motion. The "peak frame" is defined as the local extreme value of the corresponding angular curve (yaw for head shake, and pitch for head nod) during the head motion. Based on these peak frame points, a motion can be segmented into a set of sub-parts. An illustrative example is shown in Fig. 7, in which four peak frame points are detected. These four peak frames segment the motion into 3 sub-parts, that is $[p_1, p_2]$, $[p_2, p_3]$ and $[p_3, p_4]$. Several features are derived within each sub-part including:

(1) Gradient of peak value, which is the intensity of the head motion for each sub-part

$$d_{p_t} = \left| y_{p_t} - y_{p_{t-1}} \right| \tag{4}$$

where y_{p_i} is the angular value of frame p_i.

(2) Peak to peak velocity[3]

$$V_{p_{t-1}} = \left| V_{p_t} - V_{p_{t-1}} \right| / |p_t - p_{t-1}|. \tag{5}$$

Additionally, we also use per-frame velocity as a feature for the head motion. Mathematically for frame t, the per-frame velocity is defined as:

$$v_t = y_t - y_{t-1}. \tag{6}$$

3.5. Step 4: non-manual grammatical marker recognition

We use a sequence learning approach to model the correlations between the multi-scale, spatio-temporal features and the non-manual grammatical markers of interest within each video sequence. The sequence model is trained using the concatenation of low-level and high-level features as input.

Raised and lowered eyebrows are examples of high-level features. They play important roles in many non-manual grammatical markers. For example, the ASL markers of Yes/No questions, Focus/Topics, and Conditional/When clauses are usually associated with raised eyebrows, whereas Negation and Wh-question markings are usually associated with lowered eyebrows. Head nods and shakes are also examples of high-level features; they are modeled as sequence features based on the transformation of extracted, discrete gradient and velocity values. An illustrative example is shown in Fig. 8. The discrete features of gradient and peak to peak velocities are converted into step-shape function features. The start and end points of each step reflect the peak value's location and the length of the step reflects the time interval of the corresponding peak to peak value.

With this high-level approach for feature extraction, the temporal alignment of information is implicitly encoded. These features are incorporated into a Hidden Markov Support Vector Machine (HM-SVM) [1] learning framework, which is general and scalable, and can incorporate other high-level features encoded as sequence models.

4. Experiments and discussion

We carry out our experiments on a dataset collected from native ASL signers at Boston University by Dr. C. Neidle and her research group. The original videos we used in our experiments (for both face tracking and feature extraction) had the resolution of 640 × 480 pixels. These close-shot videos focusing on the signer's face are captured with a fixed camera. Given the area ratio of the face to the whole image, the resolution of faces is approximately 220 × 300 pixels (width by height). The size of the regions of interest (ROI) is 256 × 128 pixels.

The data set contained 90 videos, corresponding to 90 ASL utterances from a single native ASL signer.[4] The videos were linguistically annotated using SignStream® [33]: manual signs, grammatical markers, and relevant non-manual behaviors (including onsets and offsets, where relevant) are labeled, and their start and end frames are identified.

[2] This general difference is robust across signers. In this data set, the average maximum amplitude is about 10 times greater for the head shakes associated with negation than for those that occur as part of wh-questions.

[3] In our work, the dataset videos are collected with the same frame per second (fps) rate, so here we directly use the frame index to represent the time.

[4] We are currently carrying out linguistic annotations of data from several additional signers. Future work will focus on recognition of these expressions across multiple signers.

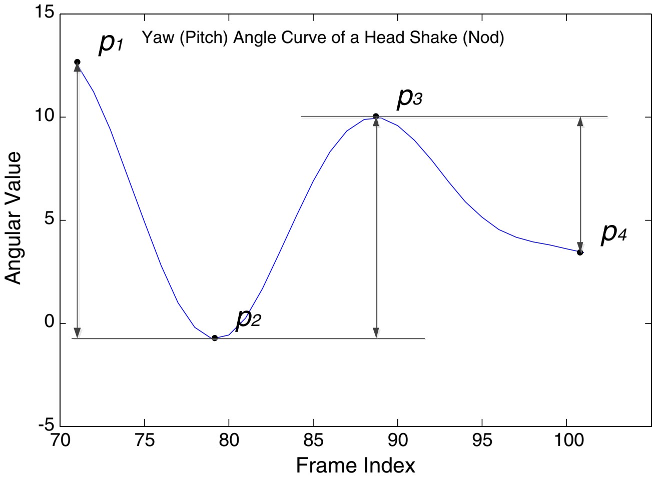

Fig. 7. Illustration of head motion high-level feature extraction.

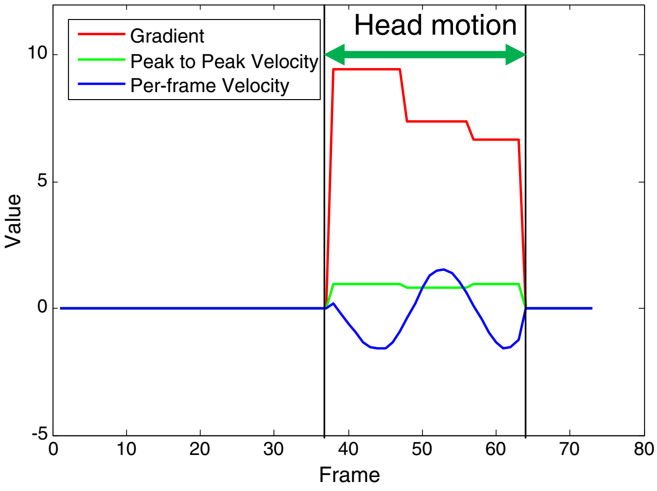

Fig. 8. Illustration of high-level feature transformation.

The size of the present data set is smaller than what was reported for our previous experiments [25]; this is because ten utterances were removed from the original dataset, for one of two reasons: (1) Because our new 3D face tracker now allows the automatic detection and reporting of tracking failures, we excluded eight videos from these experiments on recognition of non-manual events and grammatical markers. The tracking failures were caused either by serious occlusions over a long period, or by frames in which the entire face is not visible. (2) We also discarded two videos containing a non-manual grammatical marker beyond the scope of our current investigation (rhetorical question marking, which is similar in appearance to some of the non-manual markings that have been the focus of the present research). We intend to extend our research to encompass a larger set of non-manual grammatical markings in the future.

Our experiments consist of two parts. First we evaluate the results of non-manual event recognition and partitioning, the accuracy of which determines the effectiveness and robustness of the high-level features. Then we investigate the improvement in non-manual grammatical marker recognition that results from the use of the high-level features. We conduct these experiments on a Dell Workstation with a 3.4 GHz processor of eight cores and 16G memory.

4.1. Evaluation of non-manual event recognition

The non-manual event recognition is based on the use of low-level facial features. To boost the recognition accuracy for raised and lowered eyebrows, we first choose 40 video sequences for LBP feature selection: 56 phases in D_1 and 32 in D_2. The other 50 video sequences are then used for the recognition task, using a “leave one out” testing approach. This testing strategy uses one video sequence each time as a test sample while the others are used as the training set.

For raised or lowered eyebrow recognition, as mentioned previously, we use the inner, middle, and outer eyebrow height as geometric information; we calculate maxima, minima, and standard deviation of the Gabor response features from the forehead region; we reduce the LBP features from 100 to a lower dimensionality. For head shake and head nod recognition, we use yaw and pitch angles, respectively, and calculate the velocity of these two movements for each frame. In order to grasp the motion dynamics, we used a 5-frame window, which captures the dynamics of these events given a 30-frame-per second frame rate.

We introduce two measurements to assess non-manual event recognition. These measurements are based on the range of overlap between the true configuration and the detected one. The two measurements are computed as follows: $r_1 = \|R_T \cap R_D\|/\|R_T\|$ and r2 $= \|R_T \cap R_D\|/\|R_D\|$, where R_T and R_D represent the true eyebrow time range and the detected motion, respectively. r_1 evaluates whether a gesture is captured, while r_2 represents the error of over-capturing. With a threshold r_t, the recognition is considered to be valid, if $r_1 \geq r_t$ and $r_2 \geq r_t$.[5] The F1 measurement, which takes both precision and recall into account, is computed to evaluate the overall performance of the non-manual recognition: $F1_{score} = 2 \cdot precision \times recall/(precision + recall)$.

4.1.1. Accuracy of non-manual event recognition

We use the CRF Toolbox [40] to train the CRF models and test. In Table 1, we show the recognition results for raised and lowered eyebrows. We empirically set $r_t = 0.7$. The 50 video sequences contain 41 raised eyebrow and 22 lowered eyebrow events. For the recognition of raised eyebrows, we achieve an F1 score of 93.8% when only eyebrow height and Gabor response features are used. However, adding 5 dimensional LBP features improves performance: our model only fails to detect one raised eyebrow event. The performance begins to drop with a greater number of LBP features. With all of the 100 dimensional LBP features, both precision and recall significantly decrease. This is simply because more undiscriminating LBP features tend to override the important geometric and texture features, and weaken the recognition power of the CRF model.

Table 1
Recognition results for raised and lowered eyebrows.

	Raised eyebrows				Lowered eyebrows			
# LBP features	0	**5**	10	100	0	5	**10**	100
Precision	95.0%	**97.6%**	95.1%	56.6%	72.2%	73.9%	**90.9%**	48.7%
Recall	92.7%	**97.6%**	95.1%	88.7%	59.1%	77.3%	**83.3%**	57.6%
F1 Score	93.8%	**97.6%**	95.1%	69.1%	65.0%	75.6%	**86.9%**	52.7%

The recognition of lowered eyebrows is significantly worse when using only eyebrow height and Gabor response features. This is because, compared to the differences between raised and neutral eyebrows, the differences in eyebrow height between lowered and neutral eyebrows are much more subtle. With more LBP features, however, the F1 score increases; we obtain the highest score by using ten dimensional LBP features, which demonstrates that LBP features include useful information for lowered eyebrow recognition. Our experiment achieves the best performance with ten dimensional LBP features. In this case, only two lowered eyebrow events fail to be detected correctly. Both of these failure cases are due primarily to the short duration of the lowered eyebrow events.

[5] Theoretically, this criterion is slightly looser than the PASCAL criterion using intersection over union.

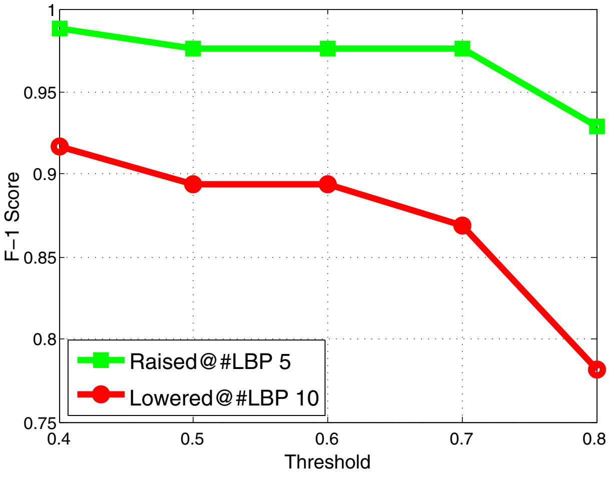

Fig. 9. Recognition results of entire raised or lowered eyebrows with different r_t.

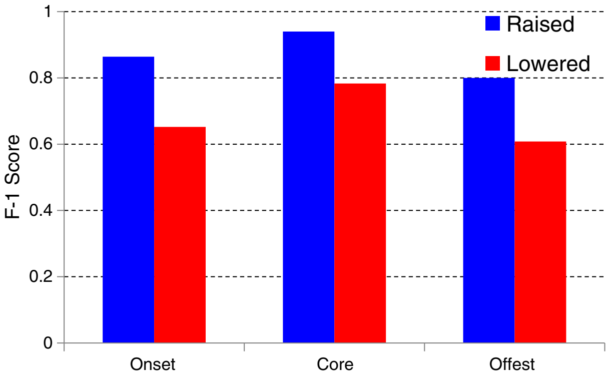

Fig. 10. F1 scores for phases of raised or lowered eyebrows with $r_t = 0.5$.

With five dimensional LBP features for recognition of raised eyebrows and ten dimensions for lowered eyebrow, we investigate the influence of r_t, i.e., the threshold value, brought into the recognition results. Fig. 9 demonstrates that both of the F1 scores drop with strict threshold value, e.g., $r_t = 0.8$. With looser r_t, the scores just slightly increase, which means we achieve high temporal accuracy. That is, the detected event overlaps to a great degree with the event as identified by the human annotators. Under similar experimental settings as above, we also evaluate the accuracy of the second-level CRFs. Fig. 10 illustrates the F1 scores for the onset, core, and offset phases with a loose threshold value. We obtain an F1 score of over 80% for the core phase of raised eyebrow, and almost 80% for the lowered eyebrows.

Table 2
Recognition results of periodic head movements.

	Head shake	Head nod
Precision	88.9%	80.0%
Recall	85.7%	80.0%
F1 Score	87.3%	80.0%

To demonstrate the effectiveness of our method for eyebrow event detection, we conducted an experimental comparison with an existing common approach. One representative approach in facial event detection is Aligned Cluster Analysis (ACA) [50], which uses unsupervised clustering. We concatenate all test videos and run the hierarchical ACA (HACA). For HACA, we first cluster the test set into twelve clusters, and then these twelve clusters are further grouped into three clusters. We assign the most probable label (raised eyebrows, lowered eyebrows, normal) to each of these three clusters. Fig. 11 shows the comparison of the recognition results for the entire eyebrow events (raised lowered). In our approach, the predicted labels are obtained by the trained CRF models. Since both the CRF model and the targeted feature potential function make full use of the label information, this method outperforms the unsupervised ACA algorithm.

Table 2 shows the recognition results for head shakes and head nods. 28 head shake events and 5 periodic head nods are included in the 50 video sequences. Missed detection emerges when a periodic head shake or head nod is very slight. We expect improvements in the future with use of larger sets of training data.

4.1.2. Temporal accuracy of non-manual event recognition

Next, we test the temporal accuracy of the motion partition achieved by the second-level CRF. Since the start point of core phase (t_s) is the beginning of the linguistically informative portion of the event, we

Table 3
Average number of frames by which the prediction of the start points of core phase differs from the annotated start point of the associated grammatical marker.

Non-manual event	# Frames
Raised eyebrows	3.6
Lowered eyebrows	4.8
Head shakes	3.1
Head nods	2.3

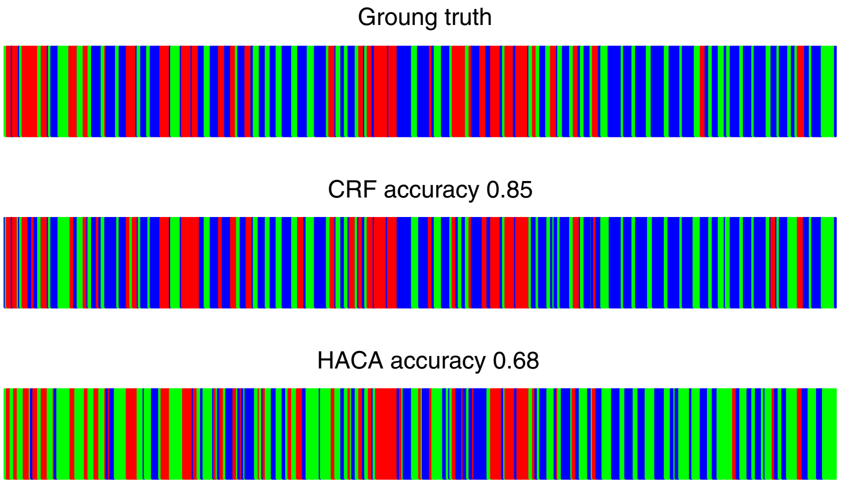

Fig. 11. Comparison results for the recognition of eyebrow events. Colors represent different eyebrow configurations (green: raised, red: lowered, blue: normal).

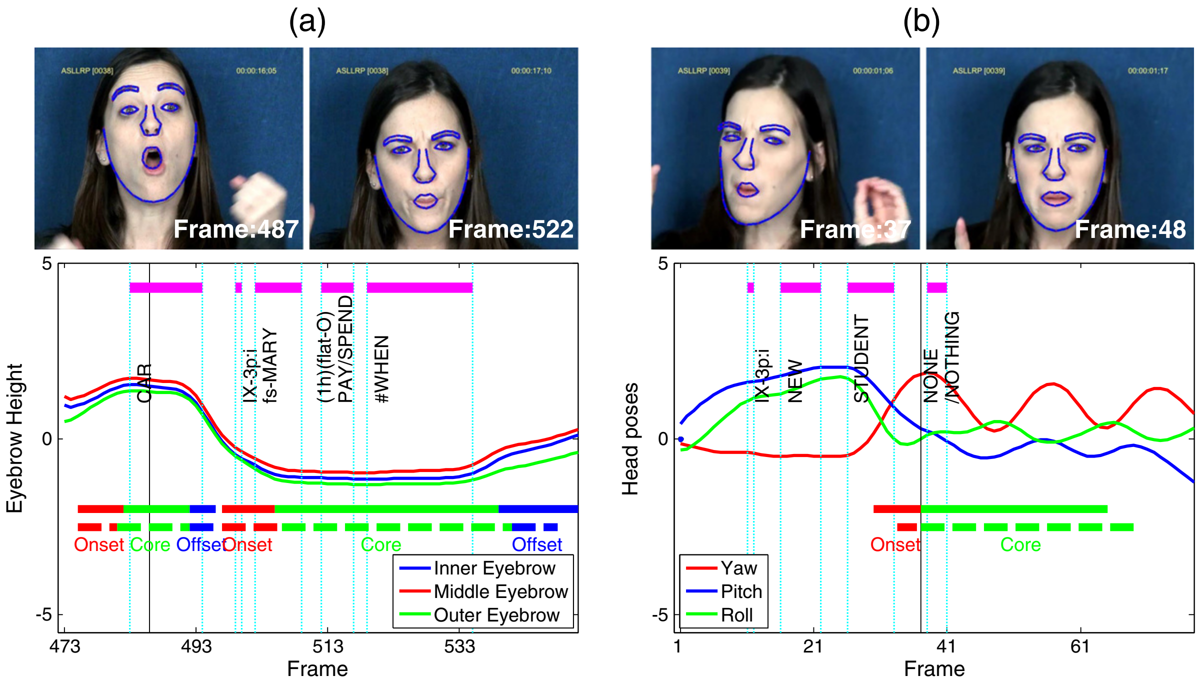

Fig. 12. Sample results for recognition of non-manual events. The solid bars represent ground truth labeling, and dashed bars show the experimental results. (a) Recognition of raised and lowered eyebrows. (b) Recognition of head shake.

evaluate the temporal accuracy of this point specifically. Especially, for head movements, if t_s^H is not at a local extremal value of yaw or pitch, we move t_s^H to the local extremum. The results are shown in Table 3. For head shakes and head nods, errors occur when t_s^H is not in the time range of the detected event based on our annotations. For eyebrow gestures, since the dynamic changes from the onset to the core phase are smooth, the boundary between the two phases is blurry, leading to less accurate localization of t_s^H. In the following, we will show that the introduction of the high-level features, based on these low-level recognition results of eyebrow gestures, further improves the grammatical marker recognition.

We show some sample results of non-manual event recognition and partitioning in Fig. 12. Fig. 12(a) contains a lowered eyebrow event following a raised eyebrow event. Our face tracker accurately locates the facial landmarks, and precise eyebrow heights are extracted. The two eyebrow gestures are correctly detected, and each is partitioned into three phases. In Fig. 12(b), we can see a yaw curve with clear periodicity, which explicitly represents a head shake. The detected start point of the core phase is very accurate.

4.2. Evaluation of grammatical marker detection

In order to verify the performance of combining the extracted low-level and high-level features in non-manual grammatical marker analysis, we conduct an experiment on detection of five specific non-manual grammatical markers in the 50 video samples we used for non-manual event recognition. The high-level features are derived based on the recognition results in Section 4.1. Table 4 shows the number of each kind of non-manual grammatical marker in our dataset. In addition to the 5 markers, "no marker (NM)" is added as another category. This indicates that none of these 5 specific markers is present, which allows us to identify cases where a non-manual marker is not present or incorrectly detected. Here we also adopt the criteria in section 4.1 to evaluate whether a detection result is valid or not.

Table 4
The number of each type of non-manual grammatical markers in our dataset.

Marker types	Sample number
Conditional/When (C/W)	14
Negation (Neg)	8
Topics/Focus (T/F)	21
Wh-question (Wh)	12
Yes/No question (Y/N)	6

We also use "leave one out" testing for non-manual grammatical marker detection evaluation. We compare our proposed low- and high-level feature combination strategy with the baseline method using only low-level features. The confusion matrices for these two approaches are shown in Table 5. The comparison between the two confusion matrices demonstrates that our combined method outperforms the low-level feature only approach in the following two respects: (1) fewer NM regions are incorrectly detected as non-manual grammatical markers; (2) our new approach more accurately recognizes and distinguishes the five types of markers. This is attributable, as discussed previously, to the fact that each type of non-manual marker is characterized by a combination of different facial expressions and head gestures, some of which are best reflected in the low-level features and others of which are associated with non-manual events that are captured by the high-level features; thus the relevant indicators have varying spatio-temporal properties.

By and large, the combined features method outperforms the method that uses only low-level features. Confusion among non-manual markers that are similar in appearance (e.g., those including raised

Table 5
Confusion matrix comparison of results obtained by using (a) both high- and low-level features (on the left), and (b) low-level features only (on the right). The label at the left of each row indicates the ground truth from the annotations.

	Both high- and low-level features					Only low-level features				
	C/W	Neg	T/F	Wh	Y/N	C/W	Neg	T/F	Wh	Y/N
C/W	**11**	0	2	0	1	**9**	0	5	1	4
Neg	0	**8**	0	0	0	0	**5**	0	2	1
T/F	1	0	**19**	0	0	5	0	**16**	2	3
Wh	0	1	0	**8**	0	1	2	0	**6**	0
Y/N	1	0	1	0	**4**	2	1	2	0	**3**
NM	0	1	2	2	0	5	6	8	7	4

Table 6
Average number of frames by which the prediction of the start frame differs from the ground truth start frame for non-manual markers that are correctly detected.

	High- and low-level features	Only low-level features
C/W	1	4
Neg	4.8	10
T/F	3.9	7
Wh	3.6	16
Y/N	3.1	7.3

brows, especially: C/W, T/F, and Y/N; or those that involve lowered brows and potentially some kind of head shake: Wh and Neg) is reduced using the new method. Nonetheless, some errors remain. We expect that the results of the new method can be further improved in the future through use of larger samples of training data and consideration of additional properties of facial expressions and head movements.

The boundaries of the non-manual events are recognized with greater temporal accuracy using the combined approach, as compared with the low-level frame-based features approach, as is evident from the improvement in identification of start points shown in Table 6. In addition, the low-level feature approach often results in identifying multiple different markers over the duration of what is really a single marker. The identification of unitary events greatly enhances the ability to demarcate individual markers. This is reflected in Table 5: whereas the sum of the values in each row for the combined method used in the left part of Table 5 corresponds to the total number of markers identified as ground truth, the sum of the rows for the low-level feature method is frequently higher because of the errors just described.

We test the improvement in the temporal accuracy of the detection results achieved by our method, by evaluating the location of the start point of the non-manual grammatical marker. Table 6 compares the average temporal accuracy using the two different feature-based methods. By separating the onset from the core part of non-manual events and encoding this as a high-level feature, we significantly improve the temporal accuracy of the non-manual grammatical marker recognition. In summary, our new approach outperforms the traditional low-level feature only methods in terms of both recognition and accuracy.

5. Conclusion

In this paper, we proposed a comprehensive framework for the automatic recognition of non-manual grammatical markers in American Sign Language (ASL). To improve the recognition accuracy, we introduced high-level features based on the detection and temporal analysis of linguistically motivated non-manual events, such as eyebrow gestures and periodic head movements. A two-level CRF is employed to identify these events, and to separate the linguistically relevant portion. We demonstrated that combining both high-level and low-level features through multi-scale, spatio-temporal analysis of head pose and face can further improve the recognition accuracy of non-manual grammatical markers in ASL, as compared with prior methods using low-level features alone. In the future, we plan to extend this method to a wider range of grammatical markers and to additional gestural components of linguistically important markings. Moreover, this approach should prove useful in differentiating other uses of eyebrow gestures from their role in signaling grammatical information in signed languages. For example, it is known that the temporal contours are different for grammatical vs. affective facial expressions that can be otherwise similar in appearance. The temporal accuracy achieved in identifying the domains of these grammatical markers is expected also to offer significant advantages for integrating information from the manual and non-manual channels (as is essential for the long-term goal of automated understanding of signed languages from video). We also intend to apply this modeling to production of more realistic sign language animations.

Acknowledgments

This project was partially supported by grants from National Science Foundation (IIS-1064965, IIS-1065013, CNS-1059281, CNS-1059218, IIS-0964597, IIS-0964385), National Space Biomedical Research Institute through NASA NCC 9-58, the Center for Identification Technology Research, a National Science Foundation (NSF) Industry/University Cooperative Research Center, and funds provided by the University of North Carolina at Charlotte. We gratefully acknowledge invaluable assistance from Rachel Benedict, Braden Painter, Joan Nash, Donna Riggle, Jessica Scott, Indya Oliver, Corbin Kuntze, Emma Preston, Tory Sampson, and many other BU students.

References

[1] Y. Altun, I. Tsochantaridis, T. Hofmann, Hidden Markov support vector machines, Proceedings of the International Conference on, Machine Learning, 2003, pp. 3–10.

[2] O. Aran, T. Burger, A. Caplier, L. Akarun, Sequential belief-based fusion of manual and non-manual information for recognizing isolated signs, Gesture-Based Human–Computer Interaction and Simulation, 2009. 134–144.

[3] C. Baker, C. Padden, Focusing on the non-manual components of American Sign Language, Understanding Language through Sign Language Research, 1978. 27–57.

[4] C. Baker-Shenk, A micro-analysis of the non-manual components of questions in American Sign Language, (Doctoral dissertation) University of California, Berkeley, CA, 1983.

[5] C. Baker-Shenk, D. Cokely, American Sign Language: A Teacher's Resource Text on Grammar and Culture, Gallaudet University Press, Washington D.C., 1980

[6] S. Chen, Y. Tian, Q. Liu, D.N. Metaxas, Recognizing expressions from face and body gesture by temporal normalized motion and appearance features, Image Vis. Comput. 31 (2) (2013) 175–185.

[7] I. Cohen, N. Sebe, A. Garg, L. Chen, T. Huang, Facial expression recognition from video sequences: temporal and static modeling, Comput. Vis. Image Underst. 91 (2003) 160–187.

[8] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 681–685.

[9] G.R. Coulter, American Sign Language typology, (Doctoral dissertation) University of California, San Diego, 1979.

[10] D. Cristinacce, T. Cootes, Feature detection and tracking with constrained local models, Proceedings of British Machine Vision Conference, vol. 3, 2006, pp. 929–938.

[11] P. Ekman, W. Friesen, Constants across cultures in the face and emotion, J. Pers. Soc. Psychol. 17 (1971) 124.

[12] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System, A Human Face, 2002.

[13] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recogn. 36 (2003) 259–275.

[14] L. Jian-zheng, Z. Zheng, Head movement recognition based on LK algorithm and gentleboost, Proceeding of IEEE International Conference on Networked Computing and Advanced Information Management, 2011, pp. 232–236.

[15] T. Joachims, Optimizing search engines using clickthrough data, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 133–142.

[16] T. Joachims, Training linear SVMs in linear time, Proceedings of ACM International Conference on Knowledge Discovery and Data Mining, 2006, pp. 217–226.

[17] A. Kanaujia, Y. Huang, D. Metaxas, Tracking facial features using mixture of point distribution models, Computer Vision, Graphics and Image Processing, Springer, 2006, pp. 492–503.

[18] D. Kelly, J. Reilly Delannoy, J. McDonald, C. Markham, Automatic Recognition of Head Movement Gestures in Sign Language Sentences, Dept. of Computer Science, National University of Ireland, 2009.

[19] M. Kendall, Rank Correlation Methods, Griffin, 1948.

[20] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, Proceedings of International Conference on, Machine Learning, 2001, pp. 282–289.

[21] C. Li, Q. Liu, J. Liu, H. Lu, Learning ordinal discriminative features for age estimation, Proceedings of IEEE Conference on Computer Vision and, Pattern Recognition, 2012, pp. 2570–2577.

[22] S. Liddell, Non-manual signals and relative clauses in American Sign Language, Understanding Language through Sign Language Research, 1978. 59–90.

[23] S.K. Liddell, American Sign Language Syntax, Mouton, The Hague, 1980.

[24] J. Lien, T. Kanade, J. Cohn, C. Li, Subtly different facial expression recognition and expression intensity estimation, Proceedings of IEEE Conference on Computer Vision and, Pattern Recognition, 1998, pp. 853–859.

[25] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. Metaxas, C. Neidle, Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL, Proceeding of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2013.

[26] D. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, C. Neidle, Recognition of non-manual markers in American Sign Language (ASL) using non-parametric adaptive 2D–3D face tracking, Proceedings of International Conference on Language Resources and Evaluation, 2012.

[27] D. Metaxas, S. Zhang, A review of motion analysis methods for human nonverbal communication computing, Image Vis. Comput. 31 (6–7) (2013) 421–433.

[28] N. Michael, D. Metaxas, C. Neidle, Spatial and temporal pyramids for grammatical expression recognition of American Sign Language, Proceedings of ACM SIGACCESS Conference on Computers and Accessibility, 2009, pp. 75–82.

[29] N. Michael, C. Neidle, D. Metaxas, Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation, 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, 2010.

[30] N. Michael, P. Yang, Q. Liu, D. Metaxas, C. Neidle, C. Center, A framework for the recognition of non-manual markers in segmented sequences of American Sign Language, Proceedings of the British Machine Vision Conference, 2011, pp. 124–131.

[31] C. Morimoto, Y. Yacoob, L. Davis, Recognition of head gestures using Hidden Markov Models, Proceedings of International Conference on Pattern Recognition, vol. 3, 1996, pp. 461–465.

[32] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 607–626.

[33] C. Neidle, SignStream^{TM} Annotation: Conventions used for the American Sign Language Linguistic Research Project, Technical Report 11, American Sign Language Linguistic Research Project Report, Boston University, 2002.

[34] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, R.G. Lee, The Syntax of American Sign Language: Functional Categories and Hierarchical Structure, MIT Press, Cambridge, MA, 2000.

[35] C. Neidle, N. Michael, J. Nash, D. Metaxas, A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus, Proceedings of ESSLLI Workshop on Formal Approaches to Sign Languages, 2009.

[36] T. Nguyen, S. Ranganath, Tracking facial features under occlusions and recognizing facial expressions in sign language, Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1–7.

[37] T. Nguyen, S. Ranganath, Recognizing continuous grammatical marker facial gestures in sign language video, Proceedings of Asian Conference on Computer Vision, 2011, pp. 665–676.

[38] O. Rudovic, V. Pavlovic, M. Pantic, Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2634–2641.

[39] S. Sarkar, B. Loeding, A.S. Parashar, Fusion of manual and non-manual information in American Sign Language recognition, Handbook of Pattern Recognition and Computer Vision, 2010.

[40] M. Schmidt, K. Swersky, Conditional Random Field (CRF) Toolbox for Matlab, 2008.

[41] C. Shan, S. Gong, P. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (2009) 803–816.

[42] T. Starner, J. Weaver, A. Pentland, Real-time American Sign Language recognition using desk and wearable computer based video, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1371–1375.

[43] Y. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 97–115.

[44] Y. Tian, T. Kanade, J. Cohn, Facial expression analysis, Handbook of Face Recognition, 2005, pp. 247–275.

[45] M. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Syst. Man Cybern. B Cybern. 42 (2012) 28–43.

[46] C. Vogler, S. Goldenstein, Facial movement analysis in ASL, Univ. Access Inf. Soc. 6 (2008) 363–374.

[47] C. Vogler, D. Metaxas, Parallel Hidden Markov Models for American Sign Language recognition, Proceedings of IEEE International Conference on Computer Vision, vol. 1, 1999, pp. 116–122.

[48] J. Wang, L. Yin, X. Wei, Y. Sun, 3d facial expression recognition based on primitive surface feature distribution, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1399–1406.

[49] H.D. Yang, S. Sclaroff, S.W. Lee, Sign language spotting with a threshold model based on Conditional Random Fields, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 1264–1277.

[50] F. Zhou, F. De la Torre, J. Cohn, Unsupervised discovery of facial events, Proceeding of IEEE Conference on Computer Vision and, Pattern Recognition, 2010, pp. 2574–2581.