Recognizing Eyebrow and Periodic Head Gestures Using CRFs for Non-Manual Grammatical Marker Detection in ASL

Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N. Metaxas and Carol Neidle

***Abstract*— Changes in eyebrow configuration, in combination with head gestures and other facial expressions, are used to signal essential grammatical information in signed languages. Motivated by the goal of improving the detection of non-manual grammatical markings in American Sign Language (ASL), we introduce a 2-level CRF method for recognition of the components of eyebrow and periodic head gestures, differentiating the linguistically significant domain (*core*) from transitional movements (which we refer to as the *onset* and *offset*). We use a robust face tracker and 3D warping to extract and combine the geometric and appearance features, as well as a feature selection method to further improve the recognition accuracy. For the second level of the CRFs, linguistic annotations were used as training for partitioning of the gestures, to separate the onset and offset. This partitioning is essential to recognition of the linguistically significant domains (in between). We then use the recognition of onset, core, and offset of these gestures together with the lower level features to detect non-manual grammatical markers in ASL.**

I. INTRODUCTION

In recent years, researchers have come to recognize the importance of head movements and facial expressions for computer-based sign language recognition (SLR). For example, such expressions have been used to aid in the automatic recognition of manual signs [1]–[3]. More importantly, these non-manual expressions occur in parallel with manual signing, conveying critical grammatical information [4]–[6]. The expression of such grammatical information is usually accomplished by a clustering of facial gestures and head movements. For instance, the marking of 'yes/no' questions typically involves raised eyebrows, but raised eyebrows are a component of many other grammatical markings as well [7] (e.g., topics, some types of questions, conditional clauses), which are distinguished by differences in facial expressions or head gestures.

Despite its linguistic importance, it is only relatively recently that sign language recognition has begun to focus on detection of the non-manual channel in signed languages [8]–[11]. Computer-based recognition of non-manual aspects of signed languages has made use of more general methods developed for facial expression recognition [12]–[14] and head pose estimation [15]. Traditional methods using non-manual expressions in ASL commonly constructed mapping functions between low-level facial features and grammatical markers [8]–[10]. However, low-level features do not take account of the temporal linguistic patterning of the relevant gestures. For this reason, it makes sense to take advantage of simple linguistic modeling to improve the recognition of non-manual grammatical markers by computer.

Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang and Dimitris Metaxas are with the Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA. {jl1322,lb507,shaoting,dnm}@cs.rutgers.edu

Carol Neidle is with the Linguistics Program (Department of Romance Studies) at Boston University, 621 Commonwealth Avenue, Boston, MA 02215, USA. carol@bu.edu

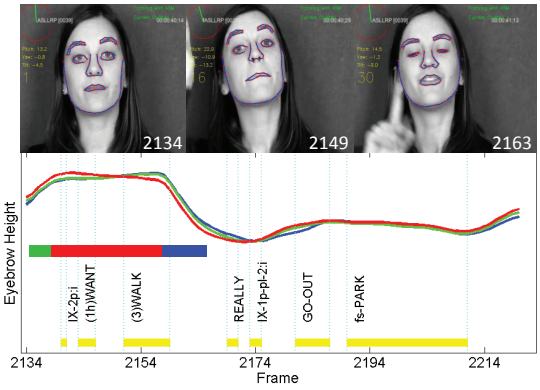


Fig. 1. In this sentence (meaning 'If you want to walk, the two of us could go out to the park,'), the first clause is marked with a typical non-manual expression for conditional modality, which includes raised eyebrows. Inner, middle, and outer eyebrow heights are shown by the blue, green, and red curves; the green, red, and blue bars identify the temporal phases of the eyebrow gesture: onset, core, and offset, respectively. The yellow bars identify the durations of the manual signs, as glossed.

This paper first focuses on the recognition of eyebrow and head gestures (specifically, only side-to-side head shakes for this research), as well as their temporal phases. We then leverage this information to improve the detection of non-manual grammatical markers in ASL. We propose a hierarchical CRF framework to automatically recognize the relevant gestures from video sequences, including their temporal phases: onset, core and offset. We use an adaptive ensemble of face trackers to get face landmarks and head poses with the Active Shape model (ASM) [16]. After that, the geometric and texture features correlated to eyebrow configurations are extracted and combined. For improved recognition results of eyebrow gestures, we also adopt a ranking method to select features that retain ordinal information. In our two-level CRF framework, entire non-manual gestures are recognized on the first-level CRFs.

The second-level CRFs are used to identify the phases of the gestures. Specifically, for eyebrow gestures, 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*) where the eyebrows raise (or lower) to the maximal extent. The *core* of the gesture—which is linguistically aligned with the manual signing—begins at that point. The

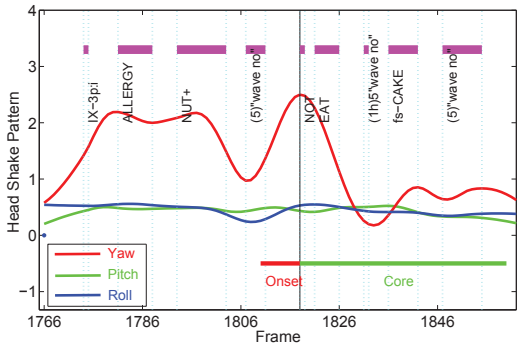


Fig. 2. Head shake pattern. The head angle *yaw* is relevant to head shake.

eyebrows tend to 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. 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 period head movements usually occurs 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. 2. After recognizing the eyebrow and head gestures, we further detect the related non-manual grammatical markers using a Hidden Markov Support Vector Machine (HM-SVM) [17] approach.

This paper is organized as follows. Section 2 describes the relevant facial features extracted through face tracking and warping, as well as the feature selection process. Section 3 introduces the hierarchical CRF framework. Section 4 presents the experiments on recognition of eyebrow and head gestures, and beyond that, on detection of grammatical markers of ASL. Section 5 contains a summary.

II. Geometric and Appearance Features

Our approach is based on a robust face tracker using deformable models, capable of estimating the 3D head motion and facial deformations due to expressions. 79 facial model points are used for tracking, and three head pose angles (i.e. pitch, roll, yaw) are estimated from the deformation of the face. After that, non-frontal faces are warped to the frontal view. Since any single type of feature would be highly sensitive and unreliable, geometric and appearance features are extracted and combined. We introduce texture features from the forehead region, which, although not in the immediate vicinity of the eyebrows, are highly relevant. These features have not been used previously for recognition of eyebrow configuration.

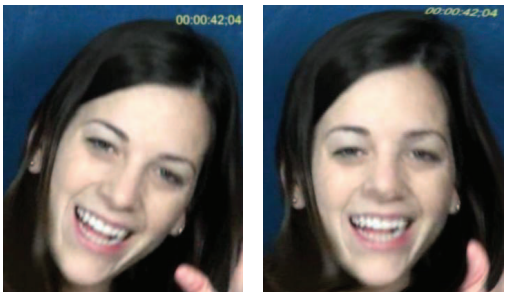
A. Face Tracking and Warping

Tracking the face for sign language recognition is more difficult than for some other applications, as a result of frequent occlusions caused by hands performing manual signs. In the traditional computer-based SLR literature, the problems with face occlusions have not been solved adequately [18] [19]. We follow the method in [9], which combined deformable models with an adaptive ensemble of face trackers using hierarchical observation likelihood. Because motion velocity and local shape deformation are taken into account, this method is capable of dealing with partial face occlusions, rapid head movements, or illumination changes. We have also adapted the sparse shape representation [20] as the facial shape priors to further improve the robustness of our system.

It is known that specific facial features captured by a fixed camera might hugely change under different head poses. For example, the aperture of eyes will appear smaller when the head pitches. Thus it is important to normalize the geometric and appearance features for varying head poses. In this paper, we adapt the method in [21], which used 3D shape fitting for face warping. Denote the 3D geometry of a face as s that contains n triple elements (x, y, z). A mean shape model of near-frontal 3D face $\overline{s}$ was trained in advance using Principle Component Analysis (PCA). Then a new face s' can be created from the mean $\overline{s}$ and eigenvectors v_i as:

$$s' = \overline{s} + \sum \beta_i v_i = \overline{s} + V \cdot \beta \tag{1}$$

where β is a parameter to control the transformation of the 3D deformable face model.



(a) Input face (b) warped face

Fig. 3. Face normalization using 3D face fitting.

To implement the shape fitting, coefficient β is explored to minimize the error between the projection of the pre-defined 3D geometric landmarks and the 2D face points caught by face tracker. Let $X^k = (x^k, y^k)$ be the kth landmark of the 2D face and P as the projection matrix. With the energy minimization approach, the energy of the kth landmark is:

$$E_k = \frac{1}{2}\|P \cdot (\overline{s}^k + V^k \cdot \beta) - X^k\|^2 \tag{2}$$

Denote $\Lambda = diag({\lambda_1}^2, {\lambda_2}^2, \cdots, {\lambda_L}^2)$, the other energy item comes from the coefficients β is:

$$E_\beta = \frac{1}{2}\beta^T \Lambda^{-1} \beta \tag{3}$$

Therefore, the total energy function includes two items: $E = \sum w_k E_k + c \cdot E_\beta$, where w_k is the weight of the kth landmark which serves to penalize texture ambiguity. c is a parameter balancing the fitting accuracy and the shape fidelity.

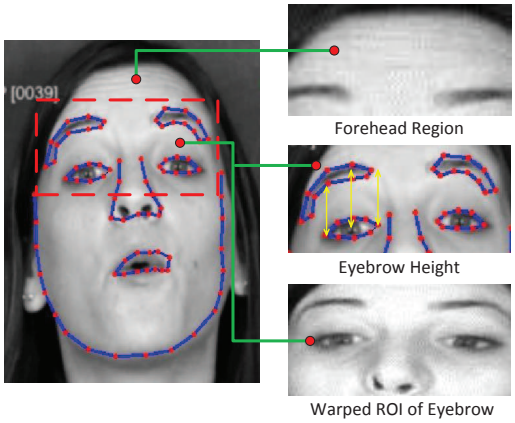


Fig. 4. Geometric and Appearance features for eyebrow gesture recognition. In the second row on the right, the three yellow lines illustrate the inner, middle and outer eyebrow height respectively, from right to left.

With the 3D face fitting, we warp all faces to frontal and extract features of pose-independence (as shown in Fig. 3), thus filtering out the effects of the head poses.

B. Feature Extraction

Identification of subtle facial expressions is a challenging task. For recognition of eyebrow gestures, selection of features that correlate with true eyebrow motions is critical for achieving accurate results. These features should be closely related to eyebrow configurations and should exhibit differences correlated with temporal phases. Intuitively, the obvious feature to describe the eyebrow gestures is eyebrow height. Geometric information is extracted from the face landmarks detected by the face tracker. The inner, middle and outer heights of the right and left eyebrows are computed.

After the face has been detected, we focus on the Region of Interest (ROI) of the eyebrows in which the Local Binary Pattern (LBP) features are extracted. Moreover, based on our observations, some facial features of the forehead are highly relevant to eyebrow gestures, yet are not contained in the ROI. As illustrated in Fig. 4, with eyebrow raise, the forehead appears more wrinkled; in contrast, lowered eyebrows are typically accompanied by brow furrowing, resulting in smooth forehead and wrinkles between the eyebrows. We calculate the Gabor responses in these two regions and use them to recognize eyebrow configurations explicitly.

C. Feature Selection

Previous work on facial expression recognition [22] [23] and age estimation [24] revealed that features or feature combinations that use rank information can further improve the performance of vision tasks involving in dynamic ordinal progress. Obviously, the eyebrow configuration undergoes ordinal changes during the onset and offset phases. During an entire eyebrow gesture, the eyebrow initially goes up/down and then goes down/up. The correct choice of features should preserve the ordinal information. Here, we only apply the feature selection process to the appearance features, since the discriminative texture features that carry ordinal information are implicit, and the redundant features should be discarded.

For feature selection, we construct our training data as follows: cut out the onset and offset phases from the entire gesture sequences, and assign ranking scores to frames according to the eyebrow heights. We divide the training data into two sets: D_1 and D_2. D_1 is used to train ranking models of individual features and D_2 to evaluate the extent to which features preserve ordinal information. For M features, we learn M ranking models using ranking SVM [25]. We then obtain the score of each feature $\{s_i\}, i = 1, \cdots, M$ on D_2. The score is calculated by the similarity between prediction list and the ground truth, as defined in [26]:

$$\tau(\widehat{L}, L) = \frac{\sum_{i,j=1}^{N} \|(\widehat{L_i} - \widehat{L_j})(L_i - L_j)\|}{N(N-1)} \quad (4)$$

where $\|\cdot\|$ is 1 if the inner function is positive and 0 when negative. $\widehat{L}$ and L are the predicted list and ground truth, respectively.

Obviously, the higher the score, the greater the ability of the feature to describe the eyebrow configurations. We rank the $s_i, i = 1, \cdots, M$ and the Kth biggest corresponding features are selected.

III. Hierarchical CRFs

Most researchers estimate eyebrow or head motion configurations directly from low-level features, using a face tracker system. However, those methods that rely on the accuracy of the tracking results are sensitive to face occlusion or abrupt head motion. Fortunately, sequence models such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) have shown advantages in event detection, sequence classification, including in the sign language domain [8] [27] [28]. 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), we utilize the CRF model to label the frames in a video sequence. As a discriminative model, the CRF model has several advantages over HMMs. It allows arbitrary dependencies between observations, and needs only a small training dataset, because there is no requirement to specify the distributions of the states and observations.

CRF is a probabilistic sequence model proposed by Lafferty et al. [29], and 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.

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

$$p(Y|X) \propto exp(\sum_{t=1}^{T}\sum_{i=1}^{N} \lambda_i f(y_t, x_t^i) + \sum_{t=1}^{T}\sum_{j=1}^{M} \mu_j g(y_t, y_{t-1}^j)) \quad (5)$$

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 $g(y_t, y_{t-1}^j)$ is the binary potential function considering the dependencies among neighborhood labels. λ_i

and μ_j are the parameters we can learn from training data using some gradient-based algorithm.

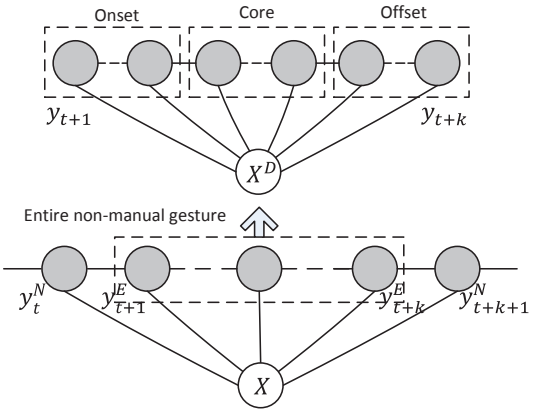


Fig. 5. The 2-level CRFs for recognition of non-manual gestures and their temporal phases. For eyebrow gestures, X are combined geometric and texture facial features; for head shake, X are the head poses. X^D refers to dynamic features using the pairwise differences.

With a given number of features, a greater number of states would lead to worse performance, because the features would not be discriminative enough to distinguish the states. Therefore, instead of detecting the three phases at one time, we accomplish the work in two stages, and each stage has comparatively fewer states. At the initial stage, what is recognized is the entirety of the non-manual gesture that involves eyebrows/head, from start to finish, without distinction among the components of the gesture, and then the further analytical decomposition of the gesture into temporal phases is done at a later stage. For this purpose, we construct a hierarchical 2-level CRF framework (Fig. 5). On the first level, the CRF models were trained to recognize entire gestures. The second level is for temporal phase identification. It should be noted that the three temporal phases have different dynamic characteristics during eyebrow raise/lower and head shake. For instance, eyebrow height increases during the onset of the eyebrow raise, while it decreases during the offset. As a result, dynamic differences among sequential features should be used in purpose. In our method, pairwise differences between neighborhood features are computed for feature construction as:

$$f_t{}' = \sum_{i=1}^{N_w} (f_{t+i} - f_{t-i}) \cdot exp(-i) \tag{6}$$

where f_t is the feature of temporal index t, N_w indicates the dynamic window size. The sign of the feature represents the direction of the dynamic change and the absolute value reflects the speed.

IV. Experiments

A. Evaluation of Eyebrow Gesture Recognition

We carried out our experiments on a dataset collected at Boston University by C. Neidle and her research group, including 100 videos, corresponding to 100 ASL utterances. The videos were linguistically annotated: manual signs and relevant non-manual behaviors (including onset and offsets, where relevant) were labeled and their start and end frames were identified. We selected 40 videos for training, and the rest were used for testing. For feature selection, we divided the training data into two parts: 20 video sequences for training ranking SVM models, and the others for evaluation. The experiments were conducted on a Dell Workstation with eight 3.4GHz processors and 16G memory.

As mentioned above, we made use of the inner, middle and outer eyebrow height as geometric information; thus we have 3 dimensional geometric features. We extract Gabor response in 2 dimensions from the forehead face region and the LBP features in 100 dimensions from the ROI of eyebrows. We reduced the dimension of LBP features to 20, 10 and 5, and then implemented our method on these feature spaces.

We introduced two measurements to assess the results of the eyebrow gesture recognition. These measurements are based on the range of overlap between the true configuration and the detected one. The two measurements, called overlap rate and pure rate, are computed as follows:

$$R_{overlap} = \frac{\|M_T \cap M_D\|}{\|M_T\|} \tag{7}$$

$$R_{pure} = \frac{\|M_T \cap M_D\|}{\|M_D\|} \tag{8}$$

where M_T and M_D represent the true eyebrow configuration and the detected motion, respectively.

$R_{overlap}$ evaluates whether a gesture is captured, while R_{pure} represents the error of over-capturing. We set a threshold $R_{threshold}$ to define the recognition results. 1) If $R_{overlap} \geq R_{threshold}$ and $R_{pure} \geq R_{threshold}$, both exit, the recognition is thought to be right. 2) $R_{overlap} < R_{threshold}$ means the gesture is not captured. 3) $R_{overlap} \geq R_{threshold}$ along with $R_{pure} < R_{threshold}$ shows that some frames are wrongly labeled by the detection result, which is regarded as a false positive.

We used the CRF Toolbox [30] to train the 2-level CRF models. We tried several different values for dynamic window size N_w, and 2 seems to be the best tradeoff between dynamic differences and local smoothness. We let $R_{threshold} = 0.4$,0.6,0.8 and conducted experiments on multiple sizes of feature spaces. The F1 measurement that takes both precision and recall into account is computed to evaluate the overall performance of the eyebrow gesture recognition, as well as the phase identification: $F1_{score} = 2 \cdot precision \times recall/(precision + recall)$.

The results are illustrated in Table I and Table II. The accuracy of the second CRFs is calculated only if an entire eyebrow gesture was correctly captured on the first level. Our method got high F1 scores on raised eyebrow recognition, even under strict criteria ($R_{threshold} = 0.8$). The best score occurs when we used 5 dimensions of LBP features. This shows that the selection of appropriate features for ROI appearance is beneficial for the recognition of eyebrow raise; however more redundant features could lead to a decline in

TABLE I

F1 SCORE OF ENTIRE EYEBROW GESTURES ON THE 1ST LEVEL CRFS.

Gestures	Feature Space	$R_{threshold}$		
		0.4	0.6	0.8
Raised	25	0.779	0.696	0.605
	15	0.831	0.765	0.612
	10	0.864	0.783	0.694
	5	0.836	0.732	0.623
Lowered	25	0.681	0.522	0.387
	15	0.732	0.614	0.447
	10	0.697	0.561	0.413
	5	0.621	0.505	0.344

TABLE II

F1 SCORE OF THE TEMPORAL PHASES ON THE SECOND LEVEL CRFS, $R_{threshold} = 0.6$

Gestures	Phases	Feature Space			
		25	15	10	5
Raised	Onset	0.165	0.394	0.650	0.544
	Core	0.821	0.826	0.918	0.897
	Offset	0.338	0.385	0.439	0.333
Lowered	Onset	0.452	0.514	0.445	0.343
	Core	0.744	0.812	0.748	0.674
	Offset	0.287	0.322	0.241	0.194

accuracy. The scores for the recognition of lowered eyebrows are comparatively worse since the lowered eyebrow configuration is closer to the neutral eyebrow expression. However, with more appearance features, the F1 score for recognition of lowered eyebrows goes up. This is because the wrinkles in the ROI caused by furrowed brows bring in more texture information, thus producing more discriminative features in this region. Hence, using comparatively more texture features would yield better results for the recognition of lowered eyebrows. Sample results for eyebrow gesture recognition are shown in Fig. 6(a); Fig. 6(b) shows an example of head shake analysis.

We also conducted the experiments on eyebrow gesture recognition using random chosen features as compared with features that preserve ordinal information; this is shown in Table III. We can see that on all the feature spaces, the selected features outperform the randomly chosen ones, illustrating that feature selection using ordinal information helps with the recognition of eyebrow gestures.

TABLE III

COMPARISON OF SELECTED FEATURES AND RANDOM FEATURES WITH F1 MEASUREMENT ON THE FIRST-LEVEL CRFS, $R_{threshold} = 0.6$

Gestures	Features	Selected Feature Space		
		20	10	5
Raised	Selected	0.696	0.765	0.783
	Random	0.435	0.630	0.689
Lowered	Selected	0.522	0.614	0.561
	Random	0.269	0.321	0.341

B. Evaluation of Detection of Grammatical Markers

We conducted experiments on the detection and classification of non-manual markings to see the extent to which the addition of the second level CRFs focused on eyebrow and periodic head gestures and their temporal phases resulted in improvements.

Here we focused on five non-manual markings: Yes/no questions; Topic or Focus; Negation; Wh-questions; and Conditional ('if') or 'when' clauses. We carried out our experiments on detecting these markers in continuous video sequences. The dataset we used here was also collected by C. Neidle and her research group; it consisted of 60 sentence-length ASL videos. 94 occurrences of non-manual grammatical markers are included (shown in Table IV).

TABLE IV

NUMBERS OF GRAMMATICAL MARKERS IN OUR DATA

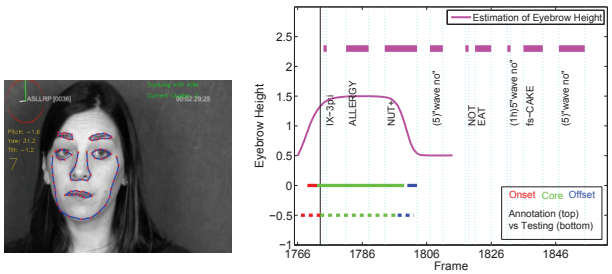
Classes	Training Samples	Testing Samples
Yes/no questions (Y/N)	4	12
Topics or Focus (Top)	23	10
Conditional/when (C/W)	14	11
Wh-questions(Whq)	5	3
Negation(Neg)	8	4

Existing methods for detection and recognition of linguistically significant non-manual markings only associate low-level features with grammatical markers directly [8] [9]. In our experiments, we use eyebrow gesture and head shake as middle-level features, and represent them as the occurrence probabilities of these non-manual expressions and of their onsets and offsets. Then we combine these middle-level features with the low-level ones, and adopt the method in [9] for detection of the grammatical markers.

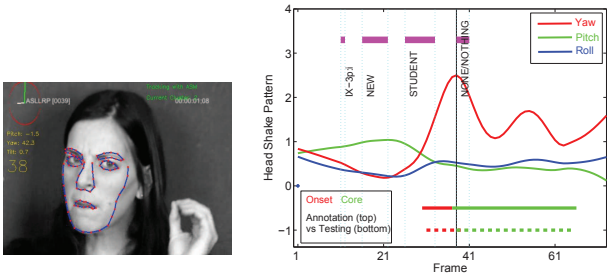
The experimental setting here is almost the same as in [9], for the purposes of comparison. We use geometric features of 105 dimensions, including eyebrow heights (inner, middle and outer), eye aperture and head poses. The 59 dimensional LBP facial features are also employed. In our method, we combine the middle-level features, i.e. recognition results of eyebrow raise, lower and head shake, with the low-level ones. For each gesture, we have the occurrence probabilities of the entire movement, its onset, core and offset. Thus the total dimension of an individual gesture is 4. With these features, the HM-SVM model is trained to detect grammatical markers in video sequences. For the baseline method, non-manual markers are detected only by low-level features. The F1 score is also used here, with $R_{threshold} = 0.5$. Since F1 score is a comprehensive measurement considering precision and recall, it can illustrate the capability of retrieving grammatical markers and avoiding wrong detections. Our method outperforms the baseline method, and achieves a 11.2% increase of the F1 score, compared to the baseline method. This is due to the fact that middle-level features provide additional spatiotemporal information that in combination with the low level features can improve the recognition of the different grammatical markers.

V. CONCLUSIONS

In this paper we propose a hierarchical CRF method to detect the linguistically relevant portion of the eyebrow or periodic head gestures. We also demonstrate that use of this technique can improve the detection of non-manual grammatical markers in ASL as compared with an approach that uses only lower level features for facial expressions.



(a) Raised eyebrow gesture



(b) Head shake gesture (Yaw depicts head shake)

Fig. 6. Some example results for recognition of eyebrow gestures and head shake, as well as their temporal phases.

Moreover, this approach should prove useful in differentiating other uses of eyebrow gestures from their role in signaling grammatical information. For example, it is known that the temporal contours are different for grammatical vs. affective facial expressions (and affective expressions may also involve eyebrow gestures). This will be explored in future research. We also plan to extend this approach to other component expressions of non-manual grammatical markings. We expect this will further improve the recognition accuracy for information conveyed non-manually in ASL.

VI. ACKNOWLEDGMENTS

The research reported here was partially funded by grants from the National Science Foundation (IIS-1064965, IIS-1065013, CNS-1059281, CNS-1059218, IIS-0964597, and IIS-0964385). We gratefully acknowledge invaluable assistance from Rachel Benedict, Braden Painter, Joan Nash, Donna Riggle, Jessica Scott, Indya Oliver, and many other BU students.

REFERENCES

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

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

[3] U. Erdem and S. Sclaroff, "Automatic detection of relevant head gestures in American Sign Language communication," in *ICPR*, vol. 1, 2002, pp. 460–463.

[4] C. Baker-Shenk and D. Cokely, *American Sign Language: A Teachers Resource Text on Grammar and Culture*. Gallaudet University Press, Washington D.C., 1980.

[5] G. Coulter, "American Sign Language Typology," Ph.D. dissertation, UCSD, 1979.

[6] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. Lee, *The syntax of American Sign Language: Functional categories and hierarchical structure*. MIT Press, 2000.

[7] C. Neidle, "SignStream^{TM} Annotation: Conventions used for the American Sign Language linguistic research project," Boston University: American Sign Language Linguistic Research Project Report, Tech. Rep. 11, 2002.

[8] T. Nguyen and S. Ranganath, "Recognizing continuous grammatical marker facial gestures in sign language video," in *ACCV*, 2011, pp. 665–676.

[9] D. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, and C. Neidle, "Recognition of nonmanual markers in American Sign Language (ASL) using non-parametric adaptive 2d-3d face tracking," in *LREC*, 2012.

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

[11] S. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *T-PAMI*, vol. 27, no. 6, pp. 873–891, 2005.

[12] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *T-PAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[13] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *PR*, vol. 36, no. 1, pp. 259–275, 2003.

[14] M. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *T-SMC B*, vol. 42, no. 1, pp. 28–43, 2012.

[15] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *T-PMAI*, vol. 31, no. 4, pp. 607 –626, 2009.

[16] T. Cootes, C. Taylor, D. Cooper, J. Graham *et al.*, "Active shape models-their training and application," *CVIU*, vol. 61, no. 1, pp. 38–59, 1995.

[17] Y. Altun, I. Tsochantaridis, T. Hofmann *et al.*, "Hidden markov support vector machines," in *Machine Learning - International Workshop then Conference-*, vol. 20, no. 1, 2003, p. 3.

[18] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *IJCV*, vol. 56, no. 3, pp. 221–255, 2004.

[19] C. Vogler and S. Goldenstein, "Facial movement analysis in ASL," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 363–374, 2008.

[20] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. Metaxas, and X. Zhou, "Sparse shape composition: A new framework for shape prior modeling," in *CVPR*, 2011, pp. 1025–1032.

[21] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3d-aware face component transfer," *TOG*, vol. 30, no. 4, p. 60, 2011.

[22] P. Yang, L. Zhong, and D. Metaxas, "Ranking model for facial age estimation," in *ICPR*, 2010, pp. 3404–3407.

[23] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *CVPR*, 2012, pp. 2634–2641.

[24] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *CVPR*, 2012, pp. 2570–2577.

[25] T. Joachims, "Training linear SVMs in linear time," in *ACM KDD*, 2006, pp. 217–226.

[26] M. Kendall, *Rank correlation methods*. Griffin, 1948.

[27] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language recognition using desk and wearable computer based video," *T-PAMI*, vol. 20, pp. 1371–1375, 1998.

[28] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *T-PAMI*, vol. 31, no. 7, pp. 1264–1277, 2009.

[29] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[30] M. Schmidt and K. Swersky, "Conditional Random Field (CRF) Toolbox for Matlab," 2008. [Online]. Available: http://www.cs.ubc.ca/~murphyk/Software/CRF/crf.html