3D Face Tracking and Multi-scale, Spatio-temporal Analysis of Linguistically Significant Facial Expressions and Head Positions in ASL

Bo Liu*, Jingjing Liu*, Xiang Yu*, Dimitris Metaxas*, Carol Neidle**

*Rutgers University, Computer Science Department,
110 Frelinghuysen Road, Piscataway, NJ 08854
**Boston University, Linguistics Program,
621 Commonwealth Ave., Boston, MA 02215
lb507@cs.rutgers.edu, jl1322@cs.rutgers.edu, xiangyu@cs.rutgers.edu, dnm@rutgers.edu, carol@bu.edu

Abstract

Essential grammatical information is conveyed in signed languages by clusters of events involving facial expressions and movements of the head and upper body. This poses a significant challenge for computer-based sign language recognition. Here, we present new methods for the recognition of nonmanual grammatical markers in American Sign Language (ASL) based on: (1) new 3D tracking methods for the estimation of 3D head pose and facial expressions to determine the relevant low-level features; (2) methods for higher-level analysis of component events (raised/lowered eyebrows, periodic head nods and head shakes) used in grammatical markings—with differentiation of temporal phases (onset, core, offset, where appropriate), analysis of their characteristic properties, and extraction of corresponding features; (3) a 2-level learning framework to combine low- and high-level features of differing spatio-temporal scales. This new approach achieves significantly better tracking and recognition results than our previous methods.

Keywords: ASL, nonmanual marker, facial expression, head position, spatio-temporal

1. Introduction

In signed languages generally, and American Sign Language (ASL) specifically, many types of important linguistic information are conveyed through facial expressions and movements of the head and upper body, which occur simultaneously with manual signing and extend over linguistic domains of varying length and duration. One particularly important function of such markings is to convey specific types of syntactic information, although the non-manual channel is also used for other linguistic and non-linguistic purposes. Computer-based sign language recognition must be able to recognize information conveyed in this way. This paper proposes a framework for automatic recognition of nonmanual grammatical markers (NMMs) in ASL based on an improved 3D face tracker and a 2-level feature extraction and classification framework. These markers convey information about the grammatical status of constituents, signaling, e.g., topics, *if/when* clauses, relative clauses, negation, and different types of questions (Baker-Shenk and Cokely, 1980; Coulter, 1979; Liddell, 1980; Neidle et al., 2000). These markings, occurring in parallel with manual signing, are built up, in part, out of component events, involving, e.g., raising or lowering of the brows and periodic head movements (nods, shakes). For example, raised eyebrows are typically found with sentence-initial topics, *if* and *when* clauses, relative clauses, and some types of questions. A head shake is essential to the nonmanual marking for negation; a head shake (albeit of smaller amplitude and greater frequency) can also mark indefiniteness, and thus is frequently present in wh-question marking.

Previous approaches to NMM recognition have generally relied on low-level features (Metaxas et al., 2012; Michael et al., 2009) or alternatively have used those low-level features to recognize higher-level gestures, such as specific types of head movements, which in turn were used for NMM detection (Nguyen and Ranganath, 2011). One significant feature of the current approach is the use of learning methods to combine low-level features – derived from a novel 3D face tracker that estimates the global head movements and facial expressions – with higher-level features of gestural events used (in varying combinations) to signal specific types of grammatical information. It is critical to separate out the preparatory phase of these high-level events, to focus on the linguistically meaningful part. For example, when raised/lowered eyebrows play a role in signaling grammatical information, the eyebrow event typically involves an "onset" phase, where the eyebrows raise/lower progressively from neutral position to attain their maximum/minimum height; the point where this occurs generally aligns with the start of the linguistic domain associated with the marking. During the final "offset" phase, the eyebrows return to neutral position. Through the proposed two-level CRF-based learning framework, these events are recognized and partitioned into the appropriate temporal phases; see Figure 1.

In order for this framework to succeed, we need accurate estimation of low-level features. The 2D trackers that have been used previously, however, have had serious tracking limitations in the presence of large head movements and occlusions. In this paper, we address these limitations through the use of a novel 3D face tracker. It is now also possible to produce visualizations as in Figure 1, showing changes in eyebrow height, eye aperture, and head position in 3D, in relation to the production of manual signs. This will be invaluable for linguistic research. Although instrumental measurements are now seen as essential for spoken language phonology, comparable data have never before been available on a large scale for sign language research.

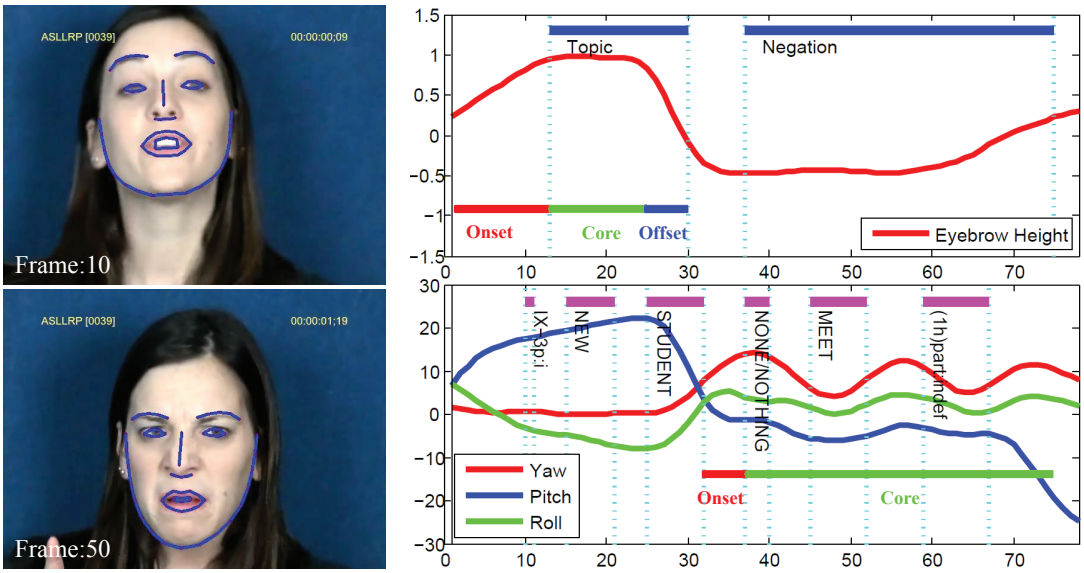


Figure 1: Eyebrow raise and head shake: "The new student, nobody has met (him/her)."

This will open up many new possibilities for linguistic research on signed languages and for cross-modality comparisons. We plan to share such data visualizations in our publicly accessible, linguistically annotated corpus (Neidle and Vogler, 2012; Neidle et al., 2014).

2. Methodology

2.1. System Overview

The framework illustrated in Figure 2 includes:

(1) Face landmark point localization and tracking using a novel 3D face model;

(2) Frame-based low-level feature extraction;

(3) Multiframe-based nonmanual event detection and partitioning; event-based high-level feature extraction;

(4) A CRF learning model trained by low- and high-level feature combination for detection and discrimination of nonmanual grammatical markers in ASL.

2.2. Facial Landmark Localization and Tracking by Optimized Part Mixtures and Cascaded Deformable Shape Model

As explained in Section 1, facial landmark localization and tracking play a fundamental role in this framework and largely determine system performance. For most current facial landmark localization and tracking algorithms, the point accuracy tends to be satisfactory for frontal faces without occlusion. However, in sign language, occlusion by the hands during sign production and large head pose changes make facial landmark extraction challenging.

Traditional landmark alignment methods, such as the Active Shape Model (ASM) (Cootes et al., 1995; Kanaujia et al., 2006), have been used to analyze ASL nonmanual markers (Michael et al., 2011; Metaxas et al., 2012; Metaxas and Zhang, 2013). Here, to improve landmark accuracy, we adopt an approach based on Optimized Part Mixtures and a Cascaded 3D Deformable Shape Model (OPM-CDSM) (Yu et al., 2013).

We use: mixtures of parts models to roughly localize the landmark points; a max-margin method to learn the weights for the landmark detector; and then a two-step cascaded deformable model to refine the landmark locations. In the first step, given each near-optimal landmark, we get the optimal alignment likelihood by searching its neighborhood. In the second step, external force constraints are used to push the landmarks to the optimal position, while shape constraints are added to preserve the shape structure. To reduce computational cost, a group sparse learning method is used to automatically select the optimized anchor points for tracking, and the selected points are reorganized into a new tree structure.

For the quantitative comparison of facial landmark localization between OPM-CDSM and other approaches, please refer to Yu et al. (2013). Overall, the 3D tracker reduces errors by at least 50% as compared to even the best 2D methods. Figure 3 compares the ASM (used in Michael et al. (2011), Metaxas et al. (2012), Liu et al. (2013), e.g.) with the OPM-CDSM in dealing with occlusions. Landmark accuracy is greatly improved with the OPM-CDSM, for the following reasons:

(1) The ASM relies mostly on local gradient search during the landmark localization, which is sensitive to many factors, including lighting conditions. It is also a 2D method dealing poorly with occlusions and 3D pose alignment.

(2) The ASM-based approach deals with head pose in discrete views, limited in number, whereas the OPM-CDSM, a 3D method, accommodates unlimited viewpoints and automatically handles alignment and 3D pose estimation.

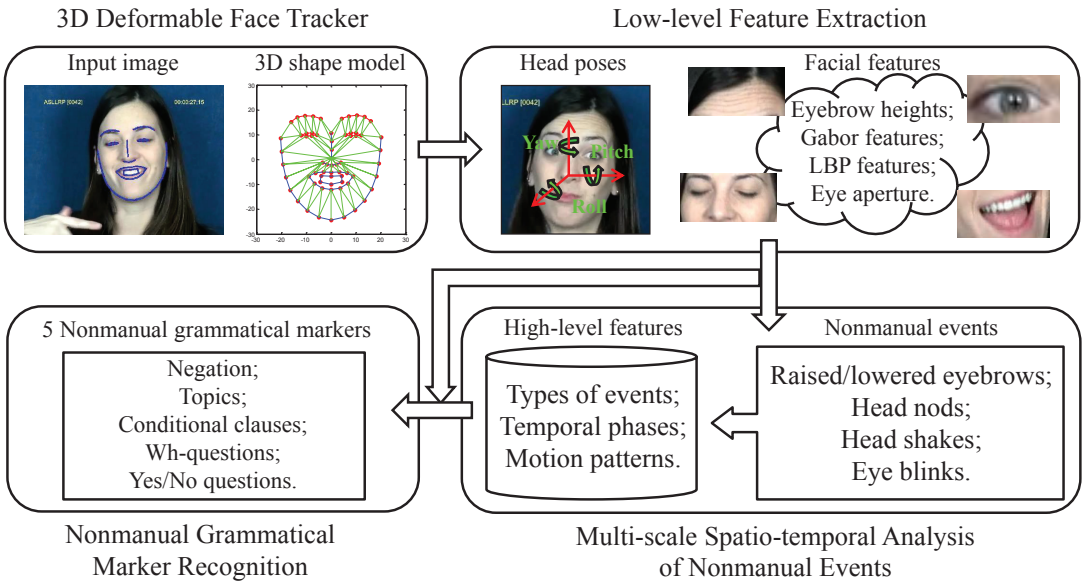


Figure 2: Flowchart of the proposed method.

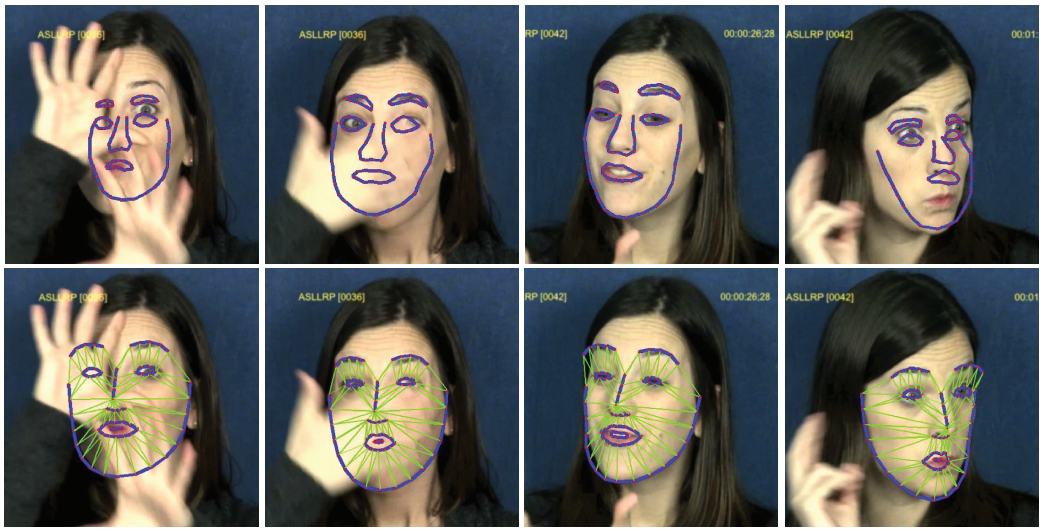


Figure 3: Comparison of the results for face landmark points using the two methods with respect to handling occlusions. The results on the top row are obtained by an ASM tracker, while the results on the bottom row are generated by our current 3D approach.

The approach outlined above allows us to estimate the multi-scale features accurately, which in turn significantly enhances NNM recognition, as discussed in Section 3.

2.3. Frame-based Low-level Feature Extraction

The comprehensive set of low-level features derived from the 3D face tracker includes:

(1) Rigid 3D head pose angles (yaw, pitch, roll), velocities, and accelerations.

(2) Appearance and Geometry features (e.g., eyebrow height, eye aperture, motion velocity and acceleration); texture features (e.g., Local Binary Pattern and Gabor response features extracted from the region of interest (ROI)).

An example is shown in Figure 4.

2.4. Event-based High-level Feature Extraction

2.4.1. NMM Event Recognition and Partitioning

NMMs often involve significant component events. The expression of many types of important grammatical information includes some combination of raised or lowered eyebrow events, where the eyebrows raise (or lower) and then remain raised (or lowered) over the relevant syntactic phrase. Similarly, periodic head movements, such as

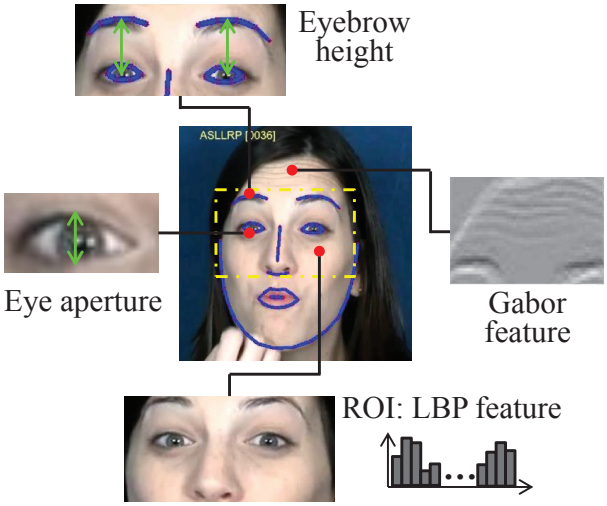


Figure 4: Illustration of low-level facial feature extraction.

nods and shakes, characterize certain NMMs. These types of events occur, with particular patterns, over domains that vary in duration. The current approach is designed to recognize these multi-scale spatio-temporal patterns so that key elements that make up NMMs can be identified and temporally localized. These types of events each have a typical anticipatory phase, where the articulators are getting into position. The beginning of the core event is generally aligned with the start point of the correlated linguistic material; furthermore, the patterning of the event, e.g., head nod or head shake, contains discriminative information for distinguishing NMMs. This motivates the current analysis of typical head and eyebrow events (beyond low-level features) relevant for NMMs.

We propose a hierarchical Conditional Random Field (CRF) framework to detect and partition the nonmanual events. CRF (Lafferty et al., 2001) is a widely used model for analysis of time series data. Given an observation sequence X, the probability of a label sequence Y has the form:

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

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 a gradient-based algorithm.

Figure 5 illustrates the framework. At the first level, CRF models are trained to recognize the entirety of the nonmanual gestures involving eyebrows or head movements, without distinguishing the gestures' components. At the second level, CRFs further analytically decompose the gesture into temporal phases.

As eyebrow movements are usually accompanied by facial texture changes (such as wrinkling of the forehead and the

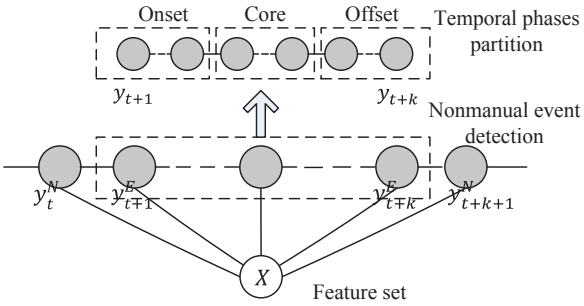


Figure 5: The 2-level CRFs for recognition of nonmanual events and their temporal phases. For eyebrow gestures, X represents the relevant low-level features of a video sequence.

area between the inner eyebrows), we also extract the texture features from these areas and combine them with eyebrow height obtained from the 3D face model. To extract the most important part from the resulting high dimensional feature vector, we adopt the Ranking SVM (Joachims, 2002) algorithm to select the most informative dimension of the texture features. For technical details please refer to Liu et al. (in press).

2.4.2. High-level Feature Extraction

Detailed motion analysis (specifically for amplitude, velocity, and frequency) is conducted—only for head shakes and head nods—based on event detection and phase partitioning. As shown in Figure 6, a given head motion can vary significantly in its patterning in different kinds of NMMs. For example, negation typically includes a head shake with a relatively large amplitude, whereas the head shake that sometimes occurs in wh-questions has a smaller movement, with more rapid repetitions. To model this difference, we develop discriminative features.

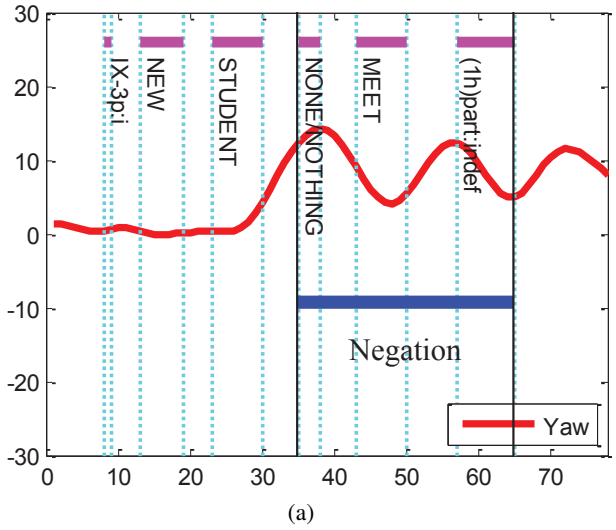
From the overall motion, we detect "peak frame" points: the local extreme values of the corresponding angular curves (yaw for head shake, pitch for head nod) during the head motion. Based on these points, the motion can be segmented. In Figure 7, the 4 detected peak frame points segment the motion into 3 parts. Several features are derived within each part, such as the gradient of peak value, peak to peak velocity, and per-frame velocity, as features for the head motion.

2.5. Combining Low-level and High-level Features for Nonmanual Grammatical Marker Recognition

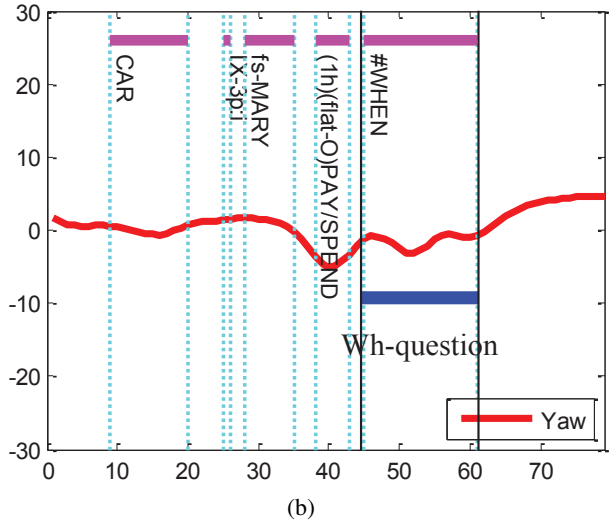
We use sequence learning to model spatio-temporal feature changes within each video sequence for NMM detection. To combine low- and high-level features, we encode the high-level features as sequence features before the feature concatenation.

3. Experiments

We conducted experiments on data collected from native ASL signers at Boston University by C. Neidle and her re-



(a)



(b)

Figure 6: Example of yaw angle curve in Negation and Wh-Question.

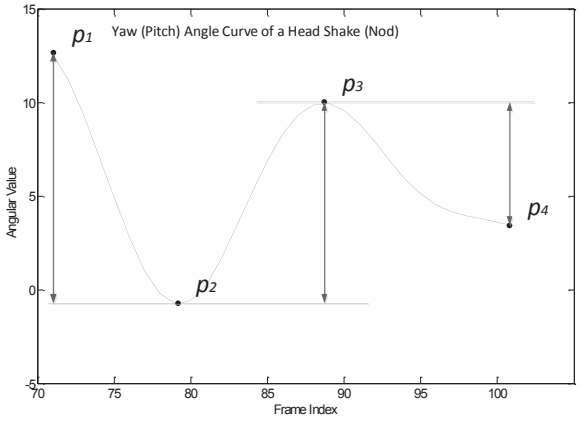


Figure 7: Illustration of head motion high-level feature extraction.

search group. The videos were linguistically annotated using SignStream® : manual signs, nonmanual events (including onsets and offsets), and NMMs were temporally localized and labelled (Neidle, 2002). Our dataset contains 149 video sequences, which is larger than the dataset we used in our previous work (Liu et al., in press), since additional videos have been collected and annotated. Each video contains between 1 and 3 NMMs. Our experiment focuses on recognizing five kinds of NMMs, namely Yes/No Questions (Y/N), Wh-Questions (Wh), Conditional/When Clauses (C/W), Negation (Neg) and Topic/Focus (Top). We conducted three experiments, to evaluate: (1) the detection of head motion and eyebrow events; (2) the performance of NMM recognition and discrimination; and (3) the accuracy of identification of the start point of NMMs. We conducted "leave one out" testing for these three experiments. We selected 85 videos from the data repository as the test set. Other videos were not used either because they contain other kinds of NMMs that we were not investigating, or because they were used for texture feature selection in eyebrow motion detection.

3.1. Experiment on Eyebrow and Head Motion Event Recognition

First we evaluated nonmanual event recognition using the selected 85 videos. These videos include different lighting conditions, partial hand occlusions, diverse head poses, and occasional motion blurrings. We successfully detected 79 raised eyebrow events of the 81 in the test dataset, 49/53 occurrences of lowered eyebrows, 38/42 head shake events, and 22/27 head nods.

3.2. Experimental Comparison of NMM Recognition

We used HM-SVM (Altun et al., 2003) for NMM recognition. The number of each kind of NMMs is shown in Table 1. We compared the NMM recognition performance obtained by using: (a) our new tracking system and a combination of low- and high-level features; (b) our new tracking system, but only low-level features; and (c) the ASM based tracker, as in Liu et al. (in press), with a combination of low- and high-level features. The confusion matrices are in Table 2.

First we compare the performance of (a) and (b). As shown in Table 2, the use of the combination of low- and high-level features results in considerable improvement in the recognition and discrimination of nonmanual grammatical markers as compared with the use of low-level features alone. Some of this improvement comes from the elimination of errors in which a single NMM was detected as multiple occurrences of NMMs. The identification of the temporal extent of the component head motion and eyebrow gestures significantly improves the delimitation of NMMs. Another benefit is the reduction of false-positive detections of NMMs, resulting from appearance and geometry features detected in individual video frames.

The comparison between Table 2(a) and Table 2(c) demonstrates that the new 3D tracking system outperforms the ASM based face tracking system for NMM recognition and results in a reduced number of false positive NMM detections. The increased accuracy of the tracking results in greater accuracy of feature extraction for both low- and high-level features, all of which contributes to the improvement in the detection and identification of NMMs.

Classes	Sample Number
Wh-Question (Wh)	7
Negation(Neg)	35
Topic/Focus (Top)	55
Yes/No Question (Y/N)	5
Conditional/When (C/W)	16

Table 1: The number of each kind of nonmanual grammatical markers in our dataset.

(a)

	Wh	Neg	Top	Y/N	C/W	NM
Wh	**6**	0	0	0	0	1
Neg	0	**34**	0	0	0	1
Top	0	0	**46**	0	6	3
Y/N	0	0	0	**5**	0	0
C/W	0	0	0	1	**15**	0
NM	1	0	1	1	0	-

(b)

	Wh	Neg	Top	Y/N	C/W	NM
Wh	**5**	1	1	0	0	1
Neg	0	**29**	3	0	0	5
Top	0	1	**41**	1	4	10
Y/N	0	1	1	**2**	1	0
C/W	0	0	1	0	**15**	0
NM	2	7	3	0	0	-

(c)

	Wh	Neg	Top	Y/N	C/W	NM
Wh	**6**	0	0	0	0	1
Neg	0	**33**	0	0	0	2
Top	0	0	**43**	0	7	5
Y/N	0	0	0	**4**	0	1
C/W	0	0	0	1	**15**	0
NM	2	3	1	1	0	-

Table 2: Confusion matrix comparison of results obtained by using (a) our new face tracking system and both low- and high- level features, (b) our new face tracking system and low-level features only, (c) the face tracking system in (Liu et al., in press) and both low- and high- level features. The label at the left of each row indicates the ground truth from the annotations: C/W (conditional or *when* clause); Neg (Negation); Top (Topic/Focus); Wh (Wh-Question); Y/N (yes/no question); NM (no marker).

3.3. Comparison of NMM Localization Accuracy

Finally we tested the improvement in temporal accuracy of NMM localization by comparing the start points of NMMs as identified through use of the combination of low- and high-level features vs. through the use of low-level features alone with the start points that human annotators had identified for those same NMMs. Although it is important to note that there is some margin of error to be expected in the human annotations, it is nonetheless apparent from Table 3 that closer agreement with the annotations about the start points is achieved through the incorporation of high-level features. This is largely attributable to the phase partitioning of the head motion and eyebrow events: the demarcation of onsets significantly improves the temporal accuracy for identification of the start points of NMMs.

	Both low- and high-level	Only low-level
Wh	2.7	7.1
Neg	3.4	7.6
Top	3.3	7.2
Y/N	2.9	4.5
C/W	1.1	3.2

Table 3: Weighted average number of frames by which the prediction of the start frame differs from the human annotation of start frame for nonmanual markers that are correctly detected through use of both low- and high-level features vs. low-level features alone.

4. Conclusion

The recognition and interpretation of the nonmanual signals that are linguistically essential in signed languages is a particularly challenging problem, as a result of the fact that the patterning of the nonmanual components occurs over varying spatio-temporal scales. In this paper, we have introduced and described a 2-level CRF learning framework for the tracking and recognition of linguistically motivated multi-scale features. We have introduced a new 3D deformable face model that achieves significantly greater accuracy in the extraction of both the low- and the high-level features used in the HM-SVM computational learning framework, which results in significantly improved detection, discrimination, and temporal localization of nonmanual grammatical markers in ASL.

Acknowledgments

The research reported here was partially funded by grants from the National Science Foundation (CNS-1059281, IIS-1064965, IIS-0964597, IIS-1065013, and CNS-0964385). We gratefully acknowledge invaluable assistance from Rachel Benedict, Braden Painter, Iryna Zhuravlova, Jessica Scott, Joan Nash, Tory Sampson, Indya Oliver, Corbin Kuntze, Amelia Wisniewski-Barker, and many other BU students.

5. References

Y. Altun, I. Tsochantaridis, and T. Hofmann. 2003. Hidden Markov Support Vector Machines. In *Proceedings of International Conference on Machine Learning*.

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

T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. 1995. Active Shape Models—Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59.

G. R. Coulter. 1979. *American Sign Language Typology*. Doctoral dissertation, University of California, San Diego.

T. Joachims. 2002. Optimizing Search Engines using Clickthrough Data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

A. Kanaujia, Y. C. Huang, and D. N. Metaxas. 2006. Tracking Facial Features using Mixture of Point Distribution Models. In *Proceedings of Indian conference on Computer Vision, Graphics and Image Processing*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of International Conference on Machine Learning*.

S. Liddell. 1980. American Sign Language Syntax. Mouton, The Hague.

J. J. Liu, B. Liu, S. T. Zhang, F. Yang, P. Yang, D. N. Metaxas, and C. Neidle. 2013. Recognizing Eyebrow and Periodic Head Gestures Using CRFs for Non-manual Grammatical Marker Detection in ASL. In *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*.

J. J. Liu, B Liu, S. T. Zhang, F. Yang, P Yang, D. N. Metaxas, and C. Neidle. in press. Non-manual Grammatical Marker Recognition based on Multi-scale, Spatio-temporal Analysis of Head Pose and Facial Expressions. *Image and Vision Computing*.

D. N. Metaxas and S. T. Zhang. 2013. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31(6-7):421–433.

D. N. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, and C. Neidle. 2012. Recognition of Nonmanual Markers in American Sign Language (ASL) using Non-Parametric Adaptive 2D-3D Face Tracking. In *Proceedings of Language Resources and Evaluation Conference*.

N. Michael, D. N. Metaxas, and C. Neidle. 2009. Spatial and Temporal Pyramids for Grammatical Expression Recognition of American Sign Language. In *Proceedings of ACM SIGACCESS Conference on Computers and Accessibility*.

N. Michael, P. Yang, Q. S. Liu, D. N. Metaxas, and C. Neidle. 2011. A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language. In *Proceedings of British Machine Vision Conference*.

C. Neidle and C. Vogler. 2012. A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface. In *Proceedings of Language Resources and Evaluation Conference*.

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

C. Neidle, J. J. Liu, B. Liu, X. Peng, C. Vogler, and D. N. Metaxas. 2014. Computer-based Tracking, Analysis, and Visualization of Linguistically Significant Non-Manual Events in American Sign Language (ASL). In *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*.

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

T. D. Nguyen and S. Ranganath. 2011. Recognizing Continuous Grammatical Marker Facial Gestures in Sign Language Video. In *Proceedings of Asian Conference on Computer Vision*.

X. Yu, J. Z. Huang, S. T. Zhang, W. Yan, and D. N. Metaxas. 2013. Pose-free Facial Landmark Fitting via Optimized Part Mixture and Cascaded Deformable Shape Model. In *Proceedings of International Conference on Computer Vision*.