Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking

Dimitris Metaxas*, Bo Liu*, Fei Yang *, Peng Yang*, Nicholas Michael*, Carol Neidle**

*Rutgers University, **Boston University
*Computer Science Department, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854
** Boston University Linguistics Program, 621 Commonwealth Ave., Boston, MA 02215
E-mail: dnm@rutgers.edu, lb507@cs.rutgers.edu, feiyang@cs.rutgers.edu,
peyang@cs.rutgers.edu, nicholam@cs.rutgers.edu, carol@bu.edu

Abstract

This paper addresses the problem of automatically recognizing linguistically significant nonmanual expressions in American Sign Language from video. We develop a fully automatic system that is able to track facial expressions and head movements, and detect and recognize facial events continuously from video. The main contributions of the proposed framework are the following: (1) We have built a stochastic and adaptive ensemble of face trackers to address factors resulting in lost face track; (2) We combine 2D and 3D deformable face models to warp input frames, thus correcting for any variation in facial appearance resulting from changes in 3D head pose; (3) We use a combination of geometric features and texture features extracted from a canonical frontal representation. The proposed new framework makes it possible to detect grammatically significant nonmanual expressions from continuous signing and to differentiate successfully among linguistically significant expressions that involve subtle differences in appearance. We present results that are based on the use of a dataset containing 330 sentences from videos that were collected and linguistically annotated at Boston University.

Keywords: linguistically based sign language recognition, model-based nonmanual expression tracking, learning-based recognition

1. Introduction

Signed languages used by the Deaf are manifested in the visual/gestural modality, but are full-fledged natural languages comparable in structure, organization, and complexity to spoken languages. Computer-based sign language recognition (SLR) offers promise for improving accessibility for the deaf, as well as for enabling efficient retrieval, annotation, and interpretation of sign language videos for research and a wide range of practical applications.

However, challenges arise from linguistic components of signed languages that occur simultaneously. In parallel with the production of signs through movements of the hands and arms, critical linguistic information is expressed in signed languages through facial expressions and head gestures that occur in complex combinations and extend over differing phrasal, scopal domains (Baker & Cokely, 1980; Coulter, 1979; Liddell, 1980; Neidle et al., 2000; Padden, 1988). These nonmanual expressions involve such things as raised or lowered eyebrows, differing degrees of eye aperture, eye gaze, nose wrinkling, head tilts and periodic head movements (nods and shakes). Such facial expressions and head gestures combine in various ways to mark the grammatical status of propositions (e.g., questions of different types, negation, conditional/when clauses, relative clauses) or the information status of constituents (e.g., topic, focus). Identification and interpretation of expressions of this kind, essential to proper interpretation of the meanings of sentences, pose significant challenges for computer-based sign language recognition (Ong & Ranganath, 2005).

In recent years, researchers have come to recognize the importance of nonmanual signals for computer-based sign language recognition (von Agris et al., 2008). Nonmanual expressions have been explored as an aid to the recognition of manual signs (Aran et al., 2009; Ming & Ranganath, 2002; Sarkar et al., 2010). Other work has focused on the component parts of grammatical expressions, such as head gestures and eyebrow movements (Erdem & Sclaroff, 2002; Kelly et al., 2009; Lafferty et al., 2001; Xu et al., 2000).

Accurate tracking of the facial landmarks is a key factor for estimation of nonmanual signals in a practical system. Although many models have been proposed for face tracking, such as Active Shape Models (Cootes et al., 1995) and Active Appearance Models (Cootes et al., 2001), most of the existing methods target global shape optimization, and are sensitive to occlusions of the face. In practice, tracking facial landmarks and recognizing facial expressions and actions is still a challenging problem, and most of the research has been limited to recognizing head gestures (Ding & Martinez, 2008) and eyebrow movements (Aran et al., 2009). However, recognizing and distinguishing the wide range of grammatical uses of subtly different combinations of such nonmanual gestures remains a difficult challenge.

Another method uses a 3D deformable model for face tracking (von Agris et al., 2008). However, this method emphasizes outlier rejection and occlusion handling at the expense of slower run time, and cannot be applied in real-time systems. A Bayesian framework and PPCA approach (Nguyen & Ranganath, 2008) have been applied to estimate real locations of occluded points. However, given a large number of parameters, the training procedure is computationally inefficient, and prone to be over-fitted to the training data. Moreover, without measuring local similarity for each tracked point, it is impossible, on that approach, to know which points are occluded, which may cause errors with respect to the occluded points to propagate to other points.

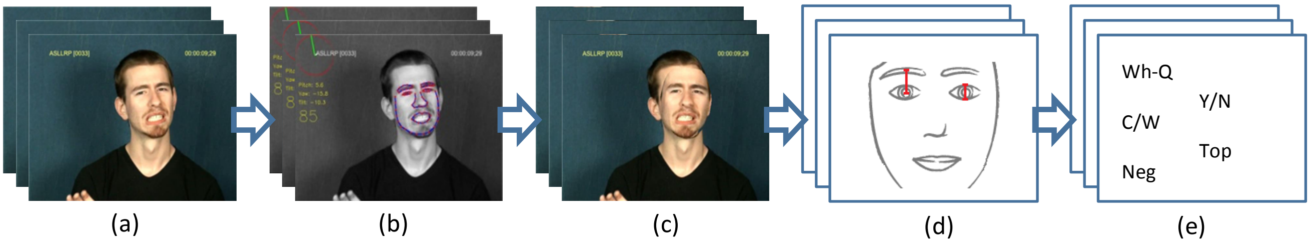

Figure 1: Overview of our system: **(a)** The input frames, **(b)** Facial landmarks being detected and tracked from the input frames, **(c)** Faces warped to frontal pose, **(d)** Geometric and appearance features extracted from the frontal faces, **(e)** Statistical sequence models applied to detect and recognize linguistically significant nonmanual markers.

Hidden Markov Models (Michael et al., 2010) and Hidden Markov Support Vector Machines (HMSVM) (Michael et al., 2011) have been applied to recognition of nonmanual markers. However, these works focus on classifying discrete markers, and are not able to recognize nonmanual markers automatically from continuous signing.

Nguyen & Ranganath (2010) detect nonmanual grammatical signals in continuous signing through the use of a 2-layer CRF model. However, this method requires manual initialization of the face tracker, thus is not able to run fully automatically. While most of the existing methods use only geometric features captured from landmark positions, we additionally use the facial texture around tracked landmarks. We argue that the geometric features are very sensitive to the accuracy of the tracking results, while the texture information is more robust in combination with the geometric features.

To overcome limitations of previous methods that do not take into account shape and texture variation in different poses, we propose a pose correction method, which uses 3D deformable shape models to warp a non-frontal face to frontal. We extract both geometric and texture information from the warped faces and build statistical sequence models to detect and recognize nonmanual signals.

In this paper, we present a new framework for tracking and analysis of facial expressions and head motions from continuous signing. The main contributions of the proposed framework are the following: (1) We have built a stochastic and adaptive ensemble of face trackers to address factors resulting in lost face track; (2) We combine 2D and 3D deformable face models to warp input frames, thus correcting for any variation in facial appearance resulting from changes in 3D head pose; (3) We use a combination of geometric features and texture features to extract a canonical frontal representation. The proposed new framework makes it possible to detect grammatically significant nonmanual expressions from continuous signing and to differentiate successfully among linguistically significant expressions that involve subtle differences in appearance.

2. System Overview

An overview of our system is shown in Figure 1. We first design a robust face tracker to track the landmark locations in each frame. Then we fit 3D deformable shape models to the tracked landmarks and warp the non-frontal faces to frontal. Next, both geometric features and appearance features are extracted from the warped faces. Finally, statistical sequence models are applied to detect and recognize nonmanual markers.

The details of our system are described below. First we describe the three main technical contributions of this paper. Then we present our learning-based approach for recognizing various nonmanual signals. Finally, we show our promising experimental results on the recognition of nonmanual markers.

2.1 Occlusion-Robust Stochastic Face Tracker

Tracking of the face for sign language recognition is a more challenging task than face tracking in other applications, because the faces of the signers are often partially occluded by hands performing manual signs or fingerspelling. Traditional face tracking methods (Baker & Matthews, 2004; Vogler & Goldenstein, 2008; Vogler et al., 2007) are not able to handle significant face occlusion. If parts of the shape are occluded, the unobservable landmarks cannot find a correct match; and the misplaced landmarks are projected back into the shape subspace, which leads to a distortion of the whole shape.

We develop a new method that addresses situations of lost track arising not only from occlusions, but also from abrupt movements and changes in illumination and appearance. The new method is based on a probabilistic and adaptive ensemble of face trackers. This is implemented as a modified particle filter (Isard & Blake, 1998) with adaptive observation likelihoods and a linear state transition model, factoring in translational and rotational velocity, and local shape deformation. Figures 2 and 3 illustrate the effect that occlusion by bangs had on the detection of eyebrow height using our previous methods (Michael et al., 2009), as well as the improvement in eyebrow tracking achieved through use of our new non-parametric tracker.

We use a hierarchical observation likelihood function consisting of the following components: (1) average texture match cost per landmark based on a set of training face images, (2) similarity to appearance templates (in the form of spatial pyramids of Local Binary Patterns) around key-points (i.e., eyebrows, eyes, nose), obtained in the first frame and updated online as tracking progresses, and (3) anthropometric constraints preserving the statistical geometry of the tracked facial shape across neighbouring frames. This results in significantly improved performance for face tracking.

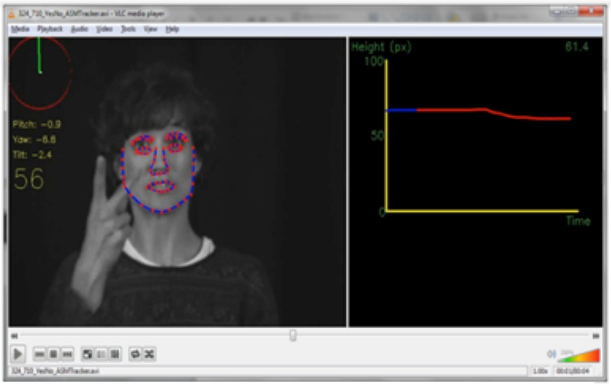

Figure 2-a: When raised eyebrows are occluded by hair, the tracker loses track and drifts downwards. It therefore incorrectly registers a drop in eyebrow height over the nonmanual topic marker (in red).

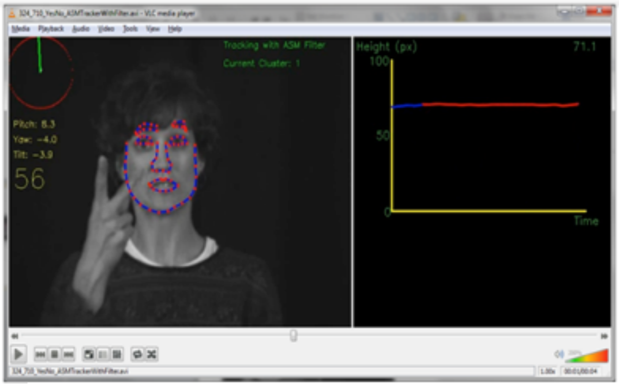

Figure 2-b: Our new non-parametric tracker correctly tracks the same sequence during the occlusion (no landmark drifting) and eyebrow height estimation is more accurate over the nonmanual marker (in red).

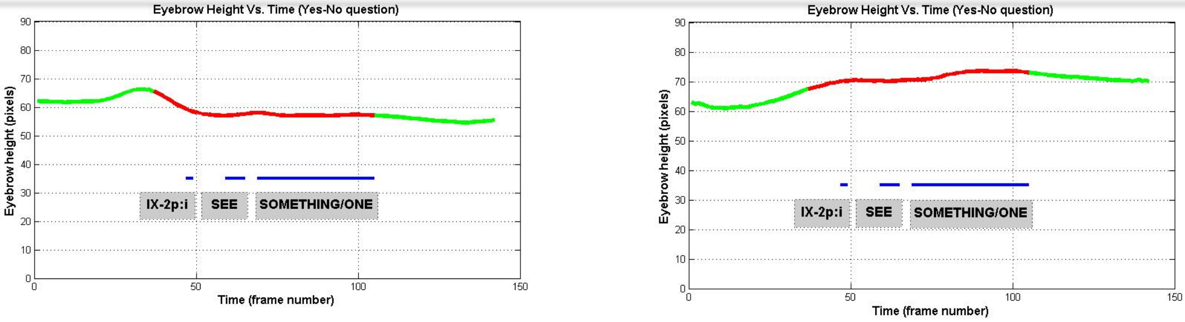

Figure 3: Plots of estimated eyebrow height over time, in relation to manual signs (represented by English-based glosses), comparing our previous 2D deterministic tracker (left) and new stochastic 2D-3D tracker (right) when eyebrow occlusion occurs (frame 35 to 140). The red line indicates yes/no question marking (frames 37-105), which characteristically includes raised eyebrows.

2.2 3D Pose Normalization

Nonmanual signals often involve frequent changes in head pose and periodic head motion (e.g., nodding, shaking). However, the geometric features and face appearances are significantly different under varying poses. Thus the characteristics of features learned in one pose cannot be directly applied to other poses. To solve this problem, we propose a pose correction method, to warp the non-frontal face region into a frontal pose.

For a given face image, we first recover its 3D pose by fitting a 3D deformable face shape. We follow the method of Yang et al. (2011), which is an efficient method for fitting 3D deformable models to near-frontal faces. The deformable face shape is defined using a shape vector s concatenating the X, Y, Z coordinates of all vertices. The deformable face model is constructed by using principal component analysis (PCA) on the training 3D shapes, and a new shape can be formed as a linear combination of eigenvectors V:

$$s = \bar{s} + V\beta$$

The 3D fitting is performed by varying the coefficients β in order to minimize the error between the projections of the pre-defined landmarks on the 3D face geometry and the 3D feature points detected by the face tracker. We use an energy minimization approach, where the total energy is defined as the total of square errors for all landmarks, and is minimized with respect to the model's projection matrix P and the model's shape coefficients β. Therefore, the projection of the kth vertex to the image plane $Y_{(k)}$ is written as:

$$Y_k = P(\bar{s}_k + V_k\beta)$$

The projection matrix P can be decomposed into scale and pose parameters. We construct a new projection matrix P′ from the same scale parameters while setting the pose parameters to be zero. With the new projection matrix, the 3D deformable face shape can be projected to the frontal pose. The displacement of each vertex forms a flow map that can be applied to warp the original non-frontal face image, as seen in Figure 4.

The proposed pose correction filters out the effects of the head orientation to the geometric features and appearance features computed from the given region. Therefore, we can learn pose-independent recognition models for nonmanual markers from only frontal faces, and apply the same model to faces of various poses. This is a significant improvement over previous methods and avoids the need to learn recognition models at various facial poses.

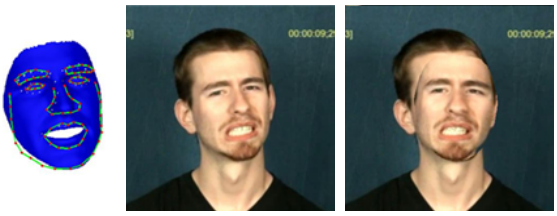

Figure 4: Pose correction. **Middle:** Input face. **Left:** The 3D shape fitted by using the landmarks detected from the input frame. **Right:** Face warped

2.3 Geometric and Appearance Features

We extract both geometric features and appearance features from the face images after 3D Pose Normalization. The geometric features are extracted from the tracked landmarks. We estimate the inner, middle and outer height of each eyebrow, and take the average for both. The speed and acceleration of the changes of heights are computed in the temporal domain. Therefore, we have a total of 9 features. Similarly, we estimate the eye heights. We also use three head pose angles (pitch, yaw and tilt, these values come from the face before 3D Pose Normalization), as well as changing speed and acceleration. Additionally, in order to capture the motion information, we employ a 5-frame window to get a patch for each frame. Namely, each frame I_t and its temporal neighbor I_{t-2}, I_{t-1}, I_{t+1} and I_{t+2} are grouped together as a patch.

Feature Description	**Dimension**
Eyebrow height	45
Eye height	15
Head pose angles	45

Table 1: The geometric features

The geometric features (eyebrow height is illustrated in Figure 3, head pose angles in Figure 6) provide measures of specific facial actions. However, they are often not able to model subtle expression changes because of noise and lighting artefacts in the image. In addition to geometric features, we also extract appearance descriptors from the regions of eyes and eyebrows that are relevant for the nonmanual markers. We extract histograms of Local Binary Patterns, which are able to model, and estimate the significant expression changes on faces.

2.4 Nonmanual Recognition based on Statistical Sequence Modeling

We apply a Hidden Markov Support Vector Machine (HMSVM) (Altun et al., 2003) to recognize nonmanual signals from video. The HMSVM models the interactions between features and class labels, as well as the interaction between neighboring labels within a sequence.

To enforce the temporal smoothness of the classified results, we apply a sliding window approach, which combines the features of all the frames inside the windows. We empirically set the window size to be 5 frames; the window is moved by one frame at a time. The above method allows the continuous recognition of markers.

3. Experiments

We first perform experiments on classifying segmented nonmanual markers. We make use of a dataset containing 330 sentences, and 406 nonmanual markers. The dataset was collected and linguistically annotated at Boston University. We randomly selected 206 sentences for training, and the remaining 124 sentences were used for testing. We focused on the nonmanual expressions associated with each of the following types of constructions, illustrated in Figure 5 (the pictures in that figure are taken from Neidle (2002), which includes more detailed descriptions):

Negation which generally includes a side-to-side headshake (usually preceded by an anticipatory head turn to the side)

Wh-questions which sometimes include a slight rapid headshake of much smaller amplitude than is found for negation (the difference is shown in Figure 5: see the red curve showing the changes in yaw angle); these are questions involving ‘who’, ‘when’, ‘what’, where’, ‘why’, ‘how’, ‘how many’, ‘which’, etc.

Yes/no questions, marking of **Topic or Focus**, and **Conditional (‘if’) or ‘when’ clauses**, which involve expressions that are subtly differentiated from one another, but normally all involve raised eyebrows.

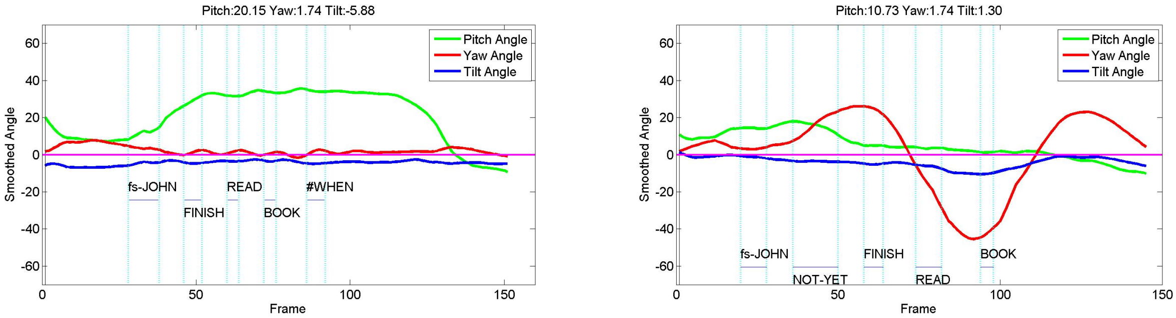

Figure 6: Geometric features extracted from a sentence: head pose angles. The yaw angle in both cases shows a periodic side-to-side headshake, as is characteristic (although different in nature) in wh-questions (left) and negation (right)

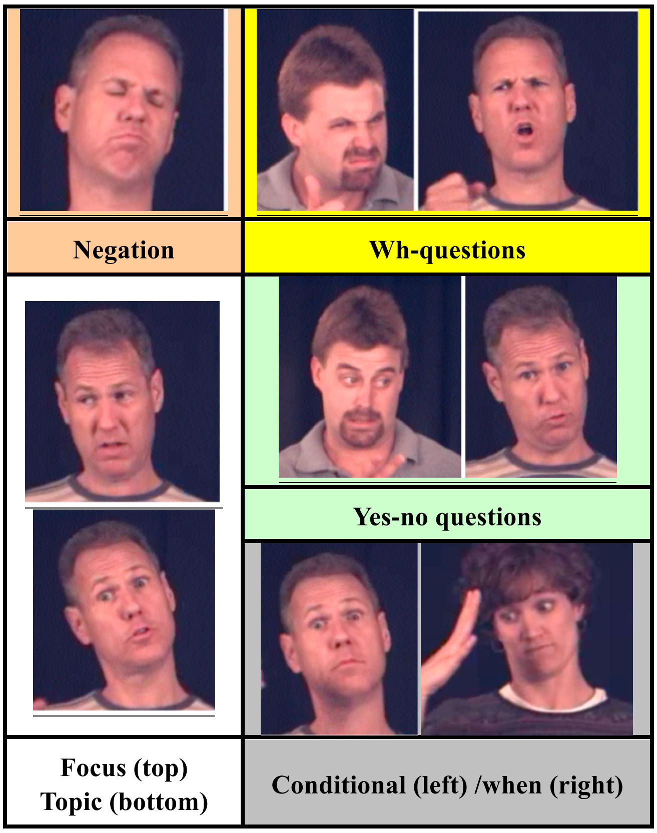

Figure 5: Typical facial expressions associated with the nonmanual markings used in this experiment

The recognition accuracy of the HMSVM model is summarized in Table 2.

Classes	Training Samples	Testing Samples
Negation (Neg)	59	24
Wh-questions (Whq)	66	40
Yes/no questions (Y/N)	32	17
Topics/focus (Top)	33	4
Conditional/when (C/W)	64	20

Table 2: Number of segmented sequences per class in our dataset

True Class	Predicted Class				
	1. Neg	2. Whq	3. Y/N	4. Top	5. C/W
1. Neg	95.8%	4.2%	0	0	0
2. Whq	2.5%	92.5%	0	2.5%	2.5%
3. Y/N	0	0	87.5%	6.3%	6.3%
4. Top	0	0	0	75.0%	25.0%
5. C/W	5.0%	0	5.0%	0	90.0%

Table 3: Confusion matrix of the proposed method

The subtleties in the distinctions among the nonmanual markings that normally include raised eyebrows make the fact that there are some confusions among markings of yes/no questions, conditionals, and topics unsurprising. Future research will be aimed at more robust discrimination of these markings. However, particularly given the close similarities among these expressions, success rates of 87-90% in recognition of yes-no questions and conditional/when clauses is particularly impressive.

We then test our system for continuous sign recognition. Sample results are shown in Figure 7. The blue lines show the ground truth; the red lines show our recognition results. We use precision and recall for

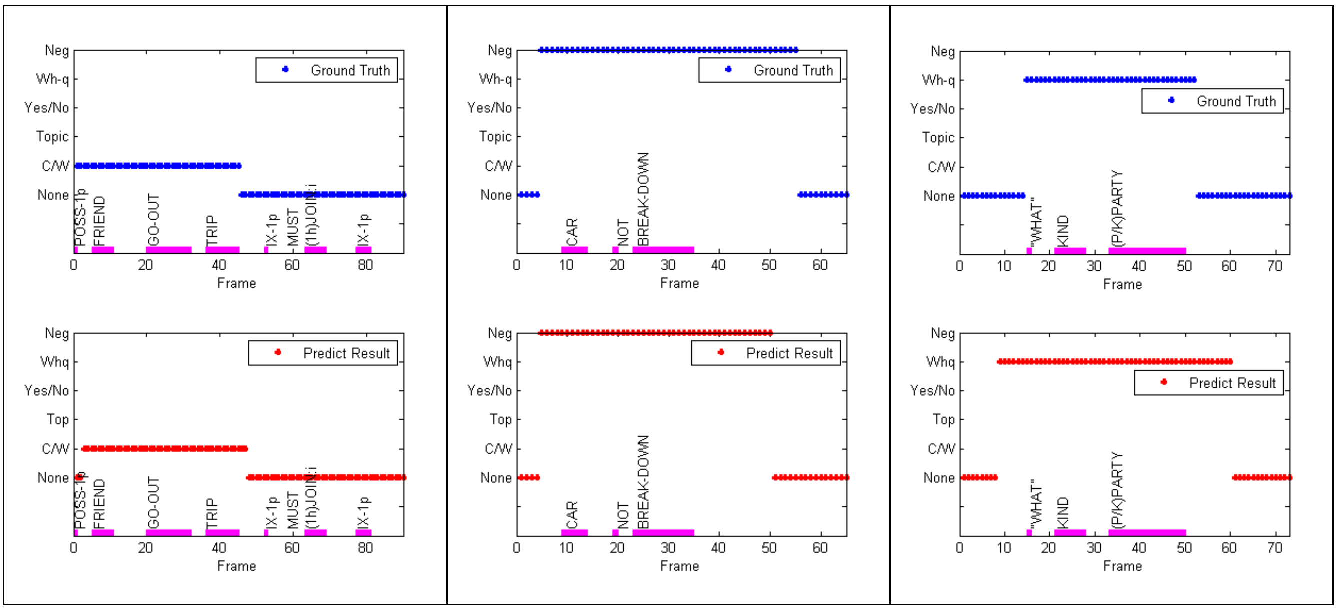

Figure 7: Performance of the proposed method in continuous sign recognition: some examples. The y-axis values correspond to the labels that were included in the annotations (from among the five listed above, or none).
The blue lines show ground truth markers; the red lines show our recognition results.
English-based glosses are provided for the manual signs, with purple lines showing the duration of each.

quantitative evaluation of the performance of continuous recognition. *Precision* is the ability to retrieve the most precise results. It is the fraction of correct detections in all detected results. For a detected marker, if over 50% of the length has the correct ground truth label, then we count it as a correct detection. *Recall* means the ability to retrieve as many elements as possible from a class. It is the fraction of successfully detected markers over all the markers in the dataset. The overall precision and recall are shown in Table 4.

Precision	81.08 %
Recall	80.37 %

Table 4: Precision and recall for continuous recognition

4. Conclusions

This paper proposes a novel framework to detect and recognize nonmanual markers of American Sign Language. Our system uses a stochastic and adaptive ensemble of face trackers, and extracts both geometric and appearance features from a canonical frontal representation. We build statistical sequence models to detect and classify nonmanual markers from continuous video input. To improve the recognition of nonmanual expressions, we used a combination of geometric features and texture features extracted from a canonical frontal representation.

The proposed new framework makes it possible to detect grammatically significant nonmanual expressions from continuous signing and to differentiate successfully among linguistically significant expressions that involve subtle differences in appearance. Based on the above innovations, our results showed, for the first time, over 80% precision and recall for the recognition of 5 types of nonmanual markers. It is our intention to improve our tracking and learning methods further to create a robust real-time nonmanual recognition system.

5. Acknowledgments

We gratefully acknowledge Braden Painter, Rachel Benedict, Norma Bowers Tourangeau, Joan Nash, Donna Riggle, Indya Oliver, and Jessica Scott for assistance with collection and annotation of the video data used here. This research has been partially funded by grants from the National Science Foundation: #CNS-0958442, #CNS-0964385, #IIS-1065013, and #IIS-0964385.

6. References

Altun, Y., I. Tsochantaridis & T. Hofmann. 2003. Hidden markov support vector machines. *International Conference on Machine Learning*.

Aran, O., T. Burger, A. Caplier & L. Akarun. 2009. Sequential Belief-Based Fusion of Manual and Non-manual Information for Recognizing Isolated Signs, Gesture-Based Human-Computer Interaction and Simulation. *LNCS (Springer) Vol. 5085*.134-44.

Baker, Charlotte & Dennis Cokely. 1980. *American Sign Language: A Teacher's Resource Text on Grammar and Culture* Silver Spring, MD: T.J. Publishers.

Baker, S. & I. Matthews. 2004. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision 56*.221-55.

Cootes, T.F., G.J. Edwards & C.J. Taylor. 2001. Active appearance models. *IEEE Trans.on Pattern Analysis and Machine Intelligence*.

Cootes, T.F., C.J. Taylor, D.H. Cooper & J. Graham. 1995. Active shape models: Their training and application. *Computer Vision and Image Understanding 61*.38-59.

Coulter, Geoffrey R. 1979. American Sign Language Typology: Doctoral Dissertation, University of California, San Diego.

Ding, L. & A.M. Martinez. 2008. Precise Detailed Detection of Faces and Facial Features. *CVPR*.

Erdem, M. & S. Sclaroff. 2002. Automatic detection of relevant head gestures in American Sign Language communication. *Proc. Int. Conf. on Pattern Recognition (ICPR)*.

Isard, M. & A. Blake. 1998. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*.

Kelly, D., J.R. Dellanoy, J. McDonald & C. Markham. 2009. Automatic Recognition of Head Movement Gestures in Sign Language Sentences.

Lafferty, J., A. McCallum & F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Paper presented at the Intl. Conference on Machine Learning.

Liddell, Scott K. 1980. *American Sign Language Syntax* The Hague: Mouton.

Michael, N., P. Yang, Q. Liu, D. Metaxas & C. Neidle. 2011. A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language. *BMVC*.

Michael, Nicholas, Dimitris Metaxas & Carol Neidle. 2009. Spatial and Temporal Pyramids for Grammatical Expression Recognition of American Sign Language. Paper presented to the 11th ACM Conference on Computers and Accessibility (ASSETS), Pittsburgh, PA, October, 2009.

Michael, Nicholas, Carol Neidle & Dimitris Metaxas. 2010. Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation. Paper presented to the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. LREC, Malta, May 22-23, 2010.

Ming, K.W. & S. Ranganath. 2002. Representations for Facial Expressions. *ICARCV*.

Neidle, Carol. 2002. *SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project.* American Sign Language Linguistic Research Project Report No. 11, Boston University.

Neidle, Carol, Judy Kegl, Dawn MacLaughlin, Benjamin Bahan & Robert G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure* Cambridge, MA: MIT Press.

Nguyen, T.D. & S. Ranganath. 2010. Recognizing Continuous Grammatical Marker Facial Gestures in Sign Language Video. *ACCV*.

Nguyen, Tan Dat & Surendra Ranganath. 2008. Tracking facial features under occlusions and recognizing facial expressions in sign language. Paper presented to the IEEE Conference on Automatic Face & Gesture Recognition, 2008.

Ong, S.C.W. & S. Ranganath. 2005. Automatic Sign Language Analysis: A survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis Intelligence 27*.873–91.

Padden, Carol A. 1988. *Interaction of Morphology and Syntax in American Sign Language* New York: Garland Publishing.

Sarkar, S. , B. Loeding & A.S. Parashar. 2010. Fusion of Manual and Non-Manual Information in American Sign Language Recognition. *Handbook of Pattern Recognition and Computer Vision (4th Edition)*, ed. by C.H. Chen. Boca Raton, FL: CRC Press.

Vogler, Christian & Siome Goldenstein. 2008. Facial movement analysis in ASL. *Universal Access in the Information Society 6*.363-74.

Vogler, Christian, Zhiguo Li, Atul Kanaujia, Siome Goldenstein & Dimitris Metaxas. 2007. The Best of Both Worlds: Combining 3D Deformable Models with Active Shape Models. Paper presented to the ICCV, 2007.

von Agris, U., M. Knorr & K. F. Kraiss. 2008. The significance of facial features for automatic sign language recognition. *International Conference on Automatic Face & Gesture Recognition.*

Xu, M., B. Raytchev, K. Sakaue, O. Hasegawa, A. Koizumi, M. Takeuchi & H. Sagawa. 2000. A Vision-Based Method for Recognizing Non-manual Information in Japanese Sign Language. *ICMI.*

Yang, F., J. Wang, E. Shechtman & L. Bourdev. 2011. Expression Flow for 3D-Aware Face Component Transfer. *ACM SIGGRAPH*.