Proceedings of the Language and Logic Workshop, Formal Approaches to Sign Languages, European Summer School in Logic, Language, and Information (ESSLLI '09), Bordeaux, France, July 20-31, 2009.

A Method for Recognition of Grammatically Significant Head Movements and Facial Expressions, Developed Through Use of a Linguistically Annotated Video Corpus[1]

CAROL NEIDLE
Boston University
carol@bu.edu

JOAN NASH
Boston University
joanpnash@gmail.com

NICHOLAS MICHAEL
Rutgers University
nicholam@cs.rutgers.edu

DIMITRIS METAXAS
Rutgers University
dnm@dragon.rutgers.edu

ABSTRACT. In Section 1 we describe a large—and expanding—set of linguistically annotated video data collected from native ASL signers. This corpus is accessible for use by the linguistics and computer science research communities and for educational purposes through a new Web-based Data Access Interface (DAI) that we have been developing to facilitate viewing, searching, and downloading relevant subsets of the data. The annotations, also available in XML format, have been carried out using SignStream®, for which a Java reimplementation with many new features (especially for efficient input of phonological information) is now underway.

[1] This research has been funded in part by grants CNS-04279883, CNS-0428231, and IIS-0705749 from the National Science Foundation. We are grateful to the students, colleagues, and linguistic consultants who have contributed to the projects described here, including especially Benjamin Bahan, Lana Cook, Quinn Duffy, Robert G. Lee, Dawn MacLaughlin, Michael Schlang, Norma Bowers Tourangeau. Also contributing to the data collection efforts at Boston University were computer scientists Stan Sclaroff, Vassilis Athitsos, and Ashwin Thangali. Development of the DAI interface has been carried out by Eugene Khafizov, Steven Li, Philip Nichols, and Ben Waber; and for SignStream® development, we are indebted to Otmar Foelsche, David Greenfield (principal programmer for version 2.2.2), and Iryna Zhuravlova (programmer for the Java reimplementation).

Proceedings of the 21st ESSLLI Language and Logic Workshop
Carlo Cecchetto and Carlo Geraci, editors,

Section 2 describes a novel method for recognition of linguistically significant head movements and facial expressions that was developed through use of our annotated corpus. Most methods for sign language recognition have dealt with tracking the hands only. However, nonmanual expressions are critical to the grammar of signed languages. Our method uses the tracked movements of the head as well as facial micro-expressions to recognize such grammatical constructions identifiable from facial expressions and head gestures. In order to track the head and estimate the head pose, we use an Active Shape Model (ASM) tracker, which fits a mask to the detected face based on a statistical model of human face appearance that is learned offline. The tracker uses texture information in the image, in the form of intensity gradients. We use a Mixture of Experts model to infer from the tracked facial landmarks the 3D pose (pitch, yaw, and tilt) of the head in every frame. Having this information, we compute in each tracked frame a bounding box around the eyes and eyebrows (whose positions are given by the tracker), and form a Region of Interest (ROI). From each ROI, we extract dense discriminative texture features in the form of pyramidal histograms of quantized SIFT descriptors. In this way, our features summarize the distribution of local texture in each frame and at different levels of detail (resolutions). For example, to identify the presence or absence of nonmanual expressions of wh-questions and negation in ASL, we use the extracted features to form a training set which is in turn used to train a Support Vector Machine to detect the presence or absence of each of these expressions. Preliminary results yield classification accuracy over 95%.

1. Linguistically annotated video corpus

In conjunction with the American Sign Language Linguistic Research Project (ASLLRP) and National Center for Sign Language and Gesture Resources (NSCLGR) at Boston University, we have been collecting an expanding set of data from native signers of American Sign Language (ASL). The data have been captured by multiple synchronized video cameras, to provide high quality video of the signing from several angles: two front (stereoscopic) views, as well as a side view and a close-up of the face. These video files are available in both compressed (30 fps) and uncompressed (60 fps) versions.

1.1 Linguistic annotation

1.1.1 Annotation software

Annotations were carried out using SignStream®, a database program designed to facilitate analysis of visual language data (MacLaughlin, Neidle, & Greenfield 2000; Neidle, Sclaroff, & Athitsos 2001; Neidle 2002b, 2003). SignStream® allows for annotation of events encoded within multiple parallel fields; events occurring in the same frame of the video are vertically aligned on the screen, as shown in Figure 1.

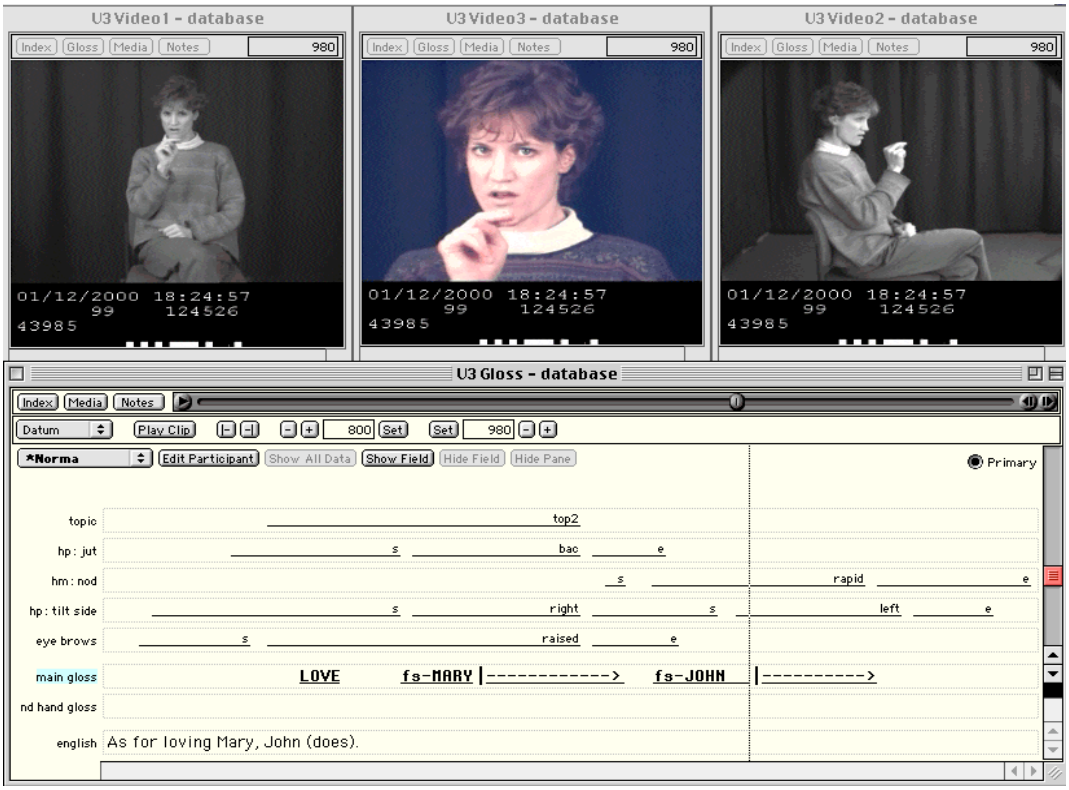

Figure 1. SignStream® version 2 - screen shot

1.1.2 Annotation conventions

Annotation conventions are documented in Neidle (2002a, 2007). In addition to identifying the start and end points of individual signs (labeled by means of conventional English glosses), the data set includes information about nonmanual behaviors, such as head nods and shakes, the position of the head and eyebrows, aperture of the eyes, expressions of the mouth, etc.. Linguistically significant clusters of gestures are also identified and labeled independently of their component parts.[2] So, for example, wh-questions generally involve both lowered eyebrows and somewhat squinted eyes, which may be accompanied by a rapid side-to-side headshake during at least part of the duration of the canonical wh-expression. Each of these components is labeled individually, but in addition, the wh-question marking itself is also labeled and associated with a start and end frame. Likewise, negation—usually involving both a slow side to side headshake and squinted eyes—is also labeled explicitly, in combination with the annotations of head movement and eye aperture. These annotations formed the basis for the research described in Section 2.

[2] Nonmanual markings associated with grammatical constructions such as negation and wh-questions have been well documented in the literature (cf. Stokoe 1960; Baker & Padden 1978; Baker 1980; Baker & Cokely 1980; Liddell 1980; Baker-Shenk 1983, 1985; Neidle, Kegl, MacLaughlin, Bahan, & Lee 2000).

1.1.3 Distribution of data files

The video and SignStream® files are distributed on a non-profit basis on CD-ROM; see the ASLLRP Web site (`http://www.bu.edu/asllrp/`) for details. However, the current version of SignStream® runs only as a Macintosh Classic application.

1.1.4 Software development: Java reimplementation of SignStream®

A new Java reimplementation is under development. The new version will be backwards-compatible and will run on multiple platforms. It will also incorporate many new features, including especially a user interface to allow efficient annotation of phonological properties for manual signs (hand shape, location, orientation, and movement, for each hand involved in the signing) from icon-based palettes. An example of the interface for data entry of hand shapes is shown in Figure 2, and the resulting screen display is illustrated in Figure 3. Once the phonological detail has been annotated for a given sign, it can be stored in the lexicon, for subsequent retrieval and insertion. Thus, this fine-grained information will not need to be encoded more than once, although adjustments to the annotations can be made to account for variations in production.

1.2 ASLLRP Data Access Interface (DAI)

The video files and linguistic annotations (also available in XML format) are also being made publicly accessible, for search and download of files that may be of interest, via a new Web interface that has just been developed: the Data Access Interface (DAI), accessible from the ASLLRP Web site: `http://www.bu.edu/asllrp/`. The DAI makes it possible to perform queries for signs or utterances based on a various criteria (singly or in combination), including characteristics of the signers, of the signs (e.g., sign type, such as fingerspelled or classifier signs; part of speech; gloss, frequency of occurrence within the database), of nonmanual expressions (e.g., raised eyebrows, wh-question marking), and/or of the video files. The video files of interest, along with corresponding annotations, can then be downloaded from the site.

1.3 Corpus development

The ASLLRP corpus currently includes fifteen spontaneous short narratives and over 400 elicited utterances. There are additional data files (including a dialogue of almost half an hour and additional elicited utterances) for which annotations are in process, and these will added to the public site as they become available.

A new collection of over 3,000 signs, produced by four native signers, is now being collected and annotated as part of a related project on large lexicon gesture representation, recognition, and retrieval: research to develop computer-based ASL search tools (Athitsos, Neidle, Sclaroff, Nash, Stefan, Yuan, & Thangali 2008). This American Sign Language Lexicon Video Dataset (ASLLVD) is also being made

accessible (`http://www.bu.edu/asllrp/lexicon/`) as new data are collected and annotated.

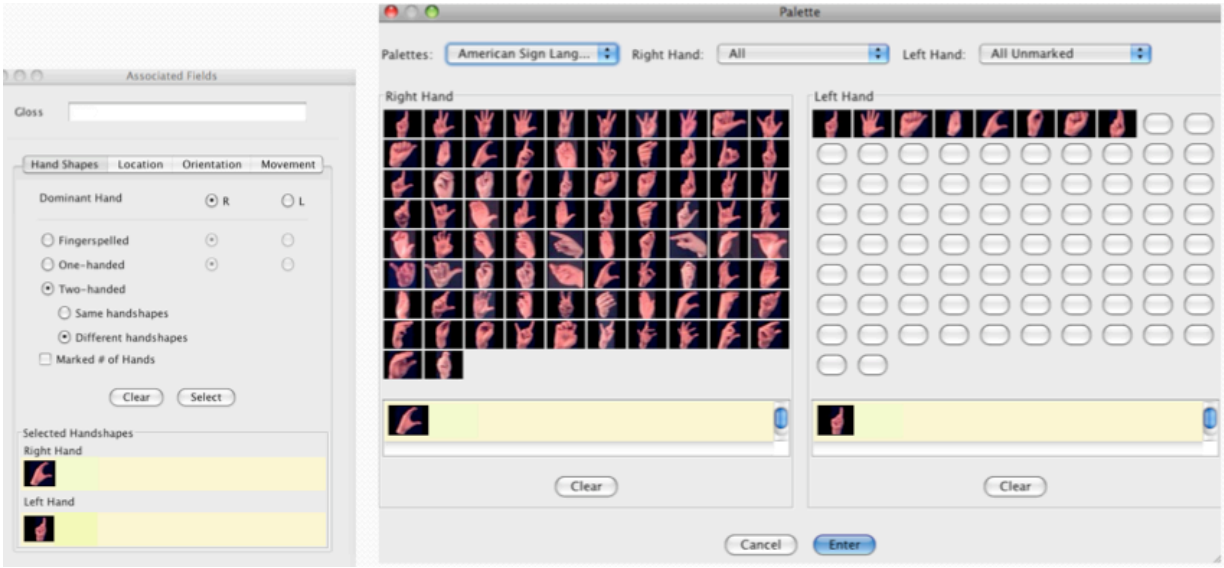

Figure 2. User interface for entry of hand shape information (SignStream® version 3)

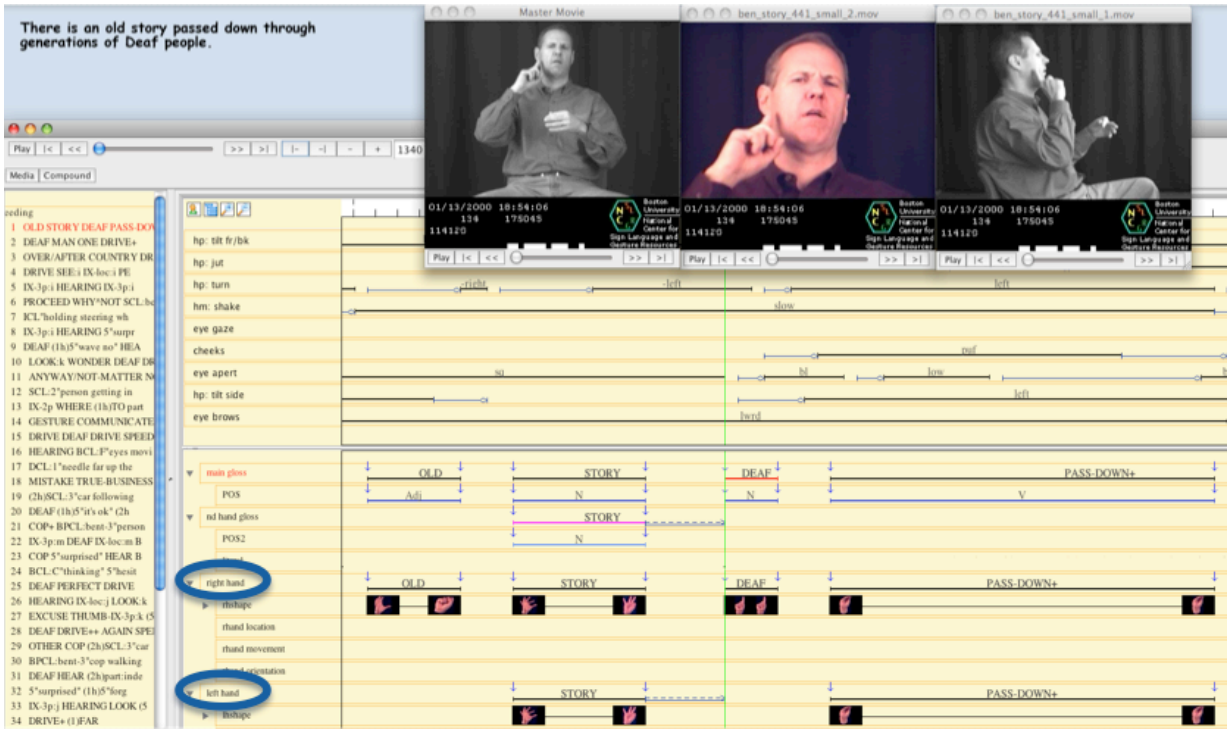

Figure 3. Sample screen display of hand shape information for dominant and non-dominant hands in SignStream® version 3.

1.4 Use of these data in computer science research

The annotated video files from the ASLLRP dataset have been used by many different researchers in linguistics and computer science. See Dreuw et al. (2008) for details of benchmark data sets based on these videos and annotations. In particular, these data have been used for tracking of ASL facial gestures (Vogler & Goldenstein 2008, e.g.). Section 2 describes the use of these data for development of capabilities for detection of linguistically significant facial expressions and head gestures.

2. Recognition of nonmanual grammatical markings

2.1 Background and Significance

In parallel with lexical items that are articulated primarily by the hands and arms, critical grammatical information is expressed nonmanually (Coulter 1979; Liddell 1980; Baker-Shenk 1983; Neidle et al. 2000, and references contained therein, e.g.). Such markings can distinguish essential syntactic properties of sentences containing otherwise identical sequences of manual signs, for example differentiating affirmative from negative sentences, declaratives from questions of different types. Thus the ability to recognize and interpret these nonmanual signals is essential to successful sign language recognition by computer. Nonetheless, with some notable exceptions (such as Vogler & Goldenstein 2008), research on computer-based sign language recognition has generally focused on detection of manual gestures (see, e.g., Bauer & Kraiss 2001; Vogler & Metaxas 2004).

2.2 Method

Our framework for facial tracking and expression recognition consists of the steps described below for each video sequence.

2.2.1 Face tracking

The video sequence is fed into an Active Shape Model (ASM) tracker to localize and track the signer's face. This tracker outputs the *(x,y)* positions of 79 facial landmarks, defining the contour of the whole face, the eyes, the eyebrows, the nose and the mouth, and the 3D head pose for each frame. Face tracking is a challenging problem because the tracker needs to generalize well to unseen faces and also handle changes in lighting. It should also cope with partial occlusions and pose changes, such as head rotations. Such difficulties cause the appearance and shape of the face to drastically change, making it harder for a computer to accurately find and track in a video.

Nevertheless, Kanaujia et al. (2006) tackle the problem with an Active Shape Model (ASM) (Cootes, Taylor, Cooper, & Graham 1995), which is a statistical model of

facial shape variation. In the ASM framework, a facial shape is represented by a number of defined landmarks, each of which is characterized by its *(x, y)* image coordinates. Through the application of Principal Component Analysis (PCA) on an aligned training set of facial shapes, it is possible to learn the major modes of shape variation within this set, essentially learning the permissible ways in which faces of different people differ.

Kanaujia et al. (2006) additionally propose learning separate ASM models for each major pose (e.g., looking frontally, left, right, up, down, etc.) and dynamically switching models as the pose of the tracked face changes through a rotation. Their system is made to run in real time by incorporating a point tracker (Shi & Tomasi 1994) to track the movement of landmarks across successive frames and only fitting the ASM model periodically to correct any tracking error. Moreover, using a Bayesian Mixture of Experts they are able to infer the 3D pose of the head in each frame from the tracked landmarks (see Kanaujia et al. (2006) for a more thorough treatment).

Figure 4. First column shows the input frame; second column the tracked face with the estimated 3D pose; third column the extracted eye and eyebrow region. The top signer is producing a wh-question expression, while the bottom signer is producing a negative expression.

The middle column of Figure 4 shows sample frames with the 79 tracked landmarks (marked as red dots), along with the predicted 3D head poses for each (shown in top left corner of each image). Pitch refers to the amount of backward tilt, yaw refers to the amount of left turn, while tilt measures the amount of clockwise head rotation. We use the tracked position of the eyes and eyebrows output by the face tracker to compute their bounding box in each frame (we will refer to the region inside this bounding box as the *eye region*). We then extract local texture features from within this bounding box to characterize the appearance of this eye region and learn to recognize lowered eyebrows and squinted eyes.

2.2.2 Feature extraction for each tracked frame utilizing ASM tracker's output

After we find the location of the bounding box for the eyes and eyebrows within a video frame, the next step is to extract useful features that can help a computer discriminate frames in which the signer's eyebrows are lowered and their eyes squinted from frames in which these gestures do not occur. To do this we use the method of the Scale Invariant Feature Transform (SIFT) by David Lowe (Lowe 2004). This is a computer vision algorithm used to detect and compute discriminative local features that can characterize the local texture and appearance of image patches. The benefit of this method is that the computed features are invariant to scale and rotation changes, meaning that they can still be detected if an image patch is resized or rotated. This is useful because as the signers are signing in ASL in a video sequence, the distance between the face and the camera may change, and there may also be rotation of the head. SIFT features allow us to maintain a good computer representation of the face's appearance, in particular around the eye region, even when the appearance of a face changes in this manner.

After we extract the discriminative SIFT features of each frame, we utilize the work on spatial pyramid representation of Lazebnik et al. (2006), which enables us to model the relationships among features within the eye regions and also provides the means for measuring feature similarity between frames. The basic idea here is that we divide the eye region into an imaginary grid of cells, covering the entire eye region, and count how many times each feature type occurs inside each cell, calling this the distribution of features within the cells. The pyramid representation creates a number of layers of such cells, but the cells in a particular layer are bigger than the cells in the layer below it. This means that even though all layers represent the same eye region, they do so at different levels of detail (i.e., different resolutions). The bottommost layer, having the smallest sized cells, forms the most detailed representation of the feature counts within an eye region, while the topmost layer forms the least detailed. Eventually, feature pyramids are represented as histograms of feature counts. Two such sample histograms are shown in Figure 5, where the blue histogram represents a frame in which the signer was producing a wh-question and the red histogram represents a frame in which the signer is producing a negative construction. Naturally, the histograms look different, especially in the lower numbered bins, which correspond to the layers of higher detail mentioned previously. The input frames together with the tracked faces and the extracted eye regions that generated these spatial pyramids are shown in Figure 4.

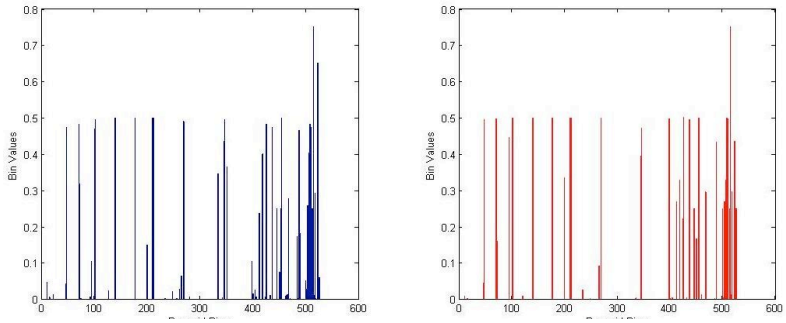

Figure 5. Sample spatial pyramids of quantized dense SIFT descriptors (50-word codebook, $\sigma=0.2$). Pyramid levels are increasing with increasing bin index. Left plot is for a wh-question expression. Right plot is for a negative expression.

To measure the similarity of two frames we just need to compare this pyramid feature representation of each frame, which basically means comparing the bins of these histograms to see the extent to which they match. Since higher levels are of a finer resolution, it is intuitive to assign a higher weight to the similarity match of cells of these levels than to that of the lower levels of coarser resolution (Lazebnik et al. 2006).

Furthermore, we propose a natural extension of this pyramid representation to the temporal domain. The ASM face tracker predicts the head pose in each frame. We compute the change in yaw angle between successive frames, which we refer to as the yaw derivative. Then we smooth these values in order to filter out any measurement noise and tracker error and construct a temporal pyramid for each video, by dividing a sequence of frames into cells, in a similar fashion as done for spatial pyramids discussed previously and using the same match strategy. In this way, we form a representation that allows us to detect the head shake of a signer. This is because we expect to see a distinctly uniform pattern of yaw angle derivatives resulting from a head shake during a negative expression, different from the pattern of yaw derivatives resulting from other ASL expressions. In other words, if there is a head shake in a video, the yaw angle will begin to increase or decrease (depending on the direction of the head shake). Soon the head will reach the peak of the turn and will need to turn back in the opposite direction, repeating this pattern for the duration of the head shake. This gives rise to the uniform pattern, since for however many frames the head turned in one direction, there should be a roughly equal number of frames turning in the other direction, too. This pattern of yaw angle derivatives during a head shake, in contrast with patterns arising from non negative expression, is illustrated in Figure 6.

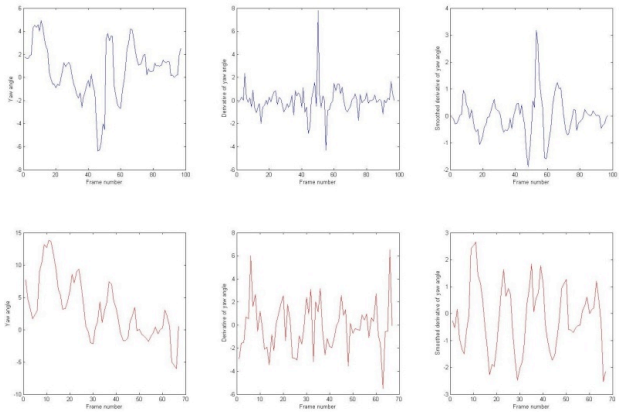

Figure 6. Sample plots of yaw angles, yaw derivatives and smoothed derivatives for two video sequences of different classes. Top row plots are from a wh-question. Bottom row plots are from a negative construction. Note the clear distinction between the patterns.

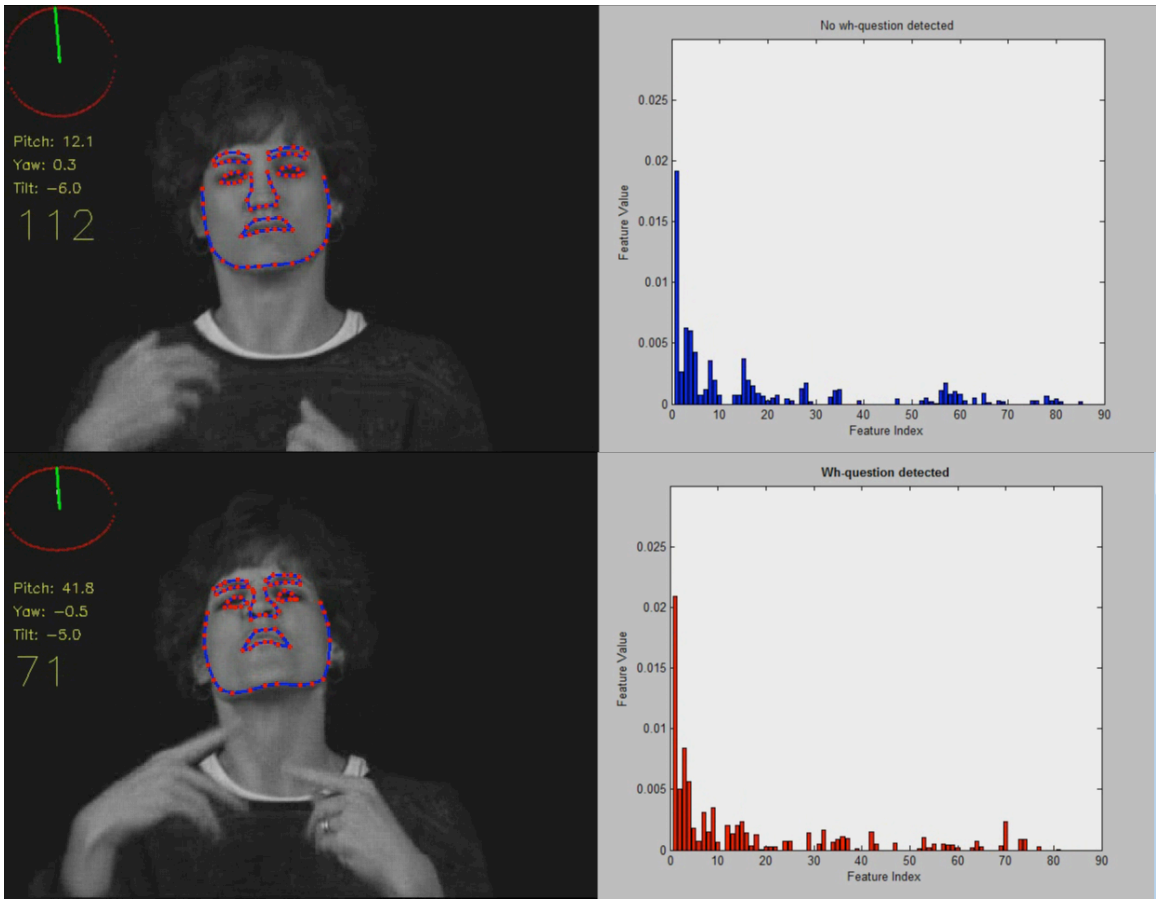

Figure 7. Sample tracked faces and extracted features
(in the bottom frame, signer is producing a wh-question expression).
Red colored features signal recognition of a wh-question expression.

2.2.3 Recognition of video sequences containing wh-questions

In order to recognize whether, in a particular video, a signer is asking a wh-question, as shown in Figure 7, we use a Support Vector Machine (SVM). This is a computer algorithm that can be trained to discriminate two classes of objects using some characteristic feature of these objects. In our case, we train an SVM to discriminate frames in which the signer's eyes appear squinted and their eyebrows appear lowered from frames in which these gestures do not occur. The discriminating feature we use is the pyramid representation of SIFT features discussed in the previous section along with the head pose, and in particular, the pitch angle.

Training is done by presenting the SVM algorithm with a set of frames in which the signer's eyes appear squinted and their eyebrows appear lowered, known as positive class examples, and a set of frames in which they are not, known as negative class examples. When presented with a video sequence to recognize, the SVM algorithm can predict, for each frame in the sequence, whether it is a wh-question frame (positive prediction) or not (negative prediction). The final decision for a video sequence is made by summing the positive predictions and the negative predictions, picking the most common prediction label within that sequence, i.e., picking the label that was predicted the most times. This method of combining the predictions of each frame to decide the final label of the video sequence is known as *majority voting*, because the majority decides the final class label.

2.2.4 Recognition of video sequences containing negative expressions

In order to recognize whether, in a particular video, a signer is producing a negative construction, we use an SVM, as in the previous section. In this case, we train an SVM to discriminate video sequences in which the signer appears to be performing a head shake from video sequences in which such a head shake does not occur. The discriminating feature we use is the pyramid representation of yaw angle derivatives.

2.3 Experimental results

The ASLLRP dataset described in Section 1 was used for this research. In our experiments, we used only the close-up view of the face. We selected a total of 47 video sequences showing utterances of wh-questions and 21 sequences showing utterances of negative expressions. These formed our set of positive examples for each of the two classes. In order to build accurate recognition models, we also needed an equal number of negative examples of each class. Therefore, we randomly selected video sequences from different classes (we refer to these as "non wh-question expression" and "non negative expression," respectively). Following this, we randomly split our two datasets of wh-questions and negative expressions into a training set and into a test set, ensuring that both sets contained data from different signers. The training sets contained about 70% of the total data and they were used to

train the algorithmic models and validate their parameters, while the remaining 30% formed the testing set, which was used to evaluate the final classifier accuracy. Table 1 shows the dataset composition in more detail.

Table 1: Dataset Composition

	Training	Testing	Total
Wh-question expression	25	22	47
Non wh-question expression	25	22	47
Total	50	44	94
Negative expression	11	10	21
Non negative expression	11	10	21
Total	22	20	42

We used the ASM face tracker of Kanaujia et al. (2006) to track the signer's face in each sequence and extract their eye region, as well as predict their 3D head pose. Figure 4 shows sample results of tracking, pose prediction, and localization of the eye region. The pose angle predictions were smoothed to filter out any measurement noise on behalf of the tracker and pose estimation algorithm. Pose angle derivatives were computed by subtracting the pose in each frame from the respective pose in the previous frame. A temporal pyramid with two levels was then constructed for each video sequence, and an SVM was trained and used to classify the test sequences into negative and non negative expressions. This achieved a recognition accuracy of 95% with only one false positive instance. Through experimentation, we found that using more levels in this temporal pyramid hurt the performance.

Similarly, from the localized eye regions we extracted SIFT features (Lowe 2004) and built spatial pyramids with 4 levels as described in Section 2.2.2. Sample spatial pyramids with four levels (i.e. $L=3$) extracted from frames in which different signers are producing different grammatical constructs, are shown in Figure 6. For recognizing the wh-questions, a single SVM classifier (which we call the "SIFT wh-question" recognizer) only achieved a recognition accuracy of 95%. We trained a second SVM (which we call the "Pose wh-question" recognizer) using the pitch angle of the signer's head in each frame as the discriminative feature, which achieved an accuracy of 73%, revealing a correlation between the pose and the expression produced by the signer in a given frame. Therefore, we implemented a stacked SVM (which we call the "Stacked wh-question recognizer") (Wolpert 1992; Tsai & Hung 2008). A stacked SVM is similar to a regular SVM, but it instead uses the outputs of other SVM's as discriminative features for performing classification, and in this way combining different features for making more accurate predictions. Our stacked SVM took as input the prediction scores output by each of the other two SVM classifiers. It classified frames into frames depicting wh-question expressions and frames depicting non wh-question expressions by combining the eye region's appearance with the pitch angle of the signer's head to make a more accurate prediction. Majority voting was

again used to decide the class label prediction of each sequence, based on the predicted labels of their constituent frames. Results are summarized in Table 2.

Table 2. Classification Accuracy

	Accuracy
Stacked wh-question	*100%*
SIFT wh-question	95%
Pose wh-question	73%
Negative expression	*95%*

2.4 Directions for future research

Currently our system uses spatial pyramids and pose angles to determine whether there is a wh-question facial expression in each frame within a video sequence. It then utilizes majority voting to do the final recognition for the video sequence. Modeling the temporal dependency between features extracted from successive frames could make the system more robust. In order to address this, we began experimenting with spatio-temporal pyramids, which also take into account the dimension of time. Such feature representation has already been used successfully to match objects in video shots (Choi, Jeon, & Lee 2008). Similarly, in order to improve our negative expression recognizer, we are also investigating methods for temporal alignment and normalization of our features, as well as richer representations, i.e., incorporating additional features. Such additions could help improve the discrimination power of our temporal pyramid representation.

3. Conclusions

As described in Section 1, the publicly available ASLLRP corpus includes a large set of linguistically annotated ASL videos. Of relevance to the research reported here are the annotations of nonmanual signals that are critical indicators of particular grammatical constructions. This dataset was used for development of techniques for recognition of such grammatically significant facial expressions and head gestures, which must be detected reliably if computer-based sign language recognition is to be successful.

This paper presents a novel framework for robust real-time face tracking and facial expression analysis from a single uncalibrated camera. Using spatial pyramids along with their temporal extension, we obtained a feature representation that is signer-independent. We demonstrated that our framework can recognize the facial components of ASL signing, by successfully recognizing wh-questions (100% accuracy) and negative expressions (95% accuracy). Lastly, we discovered a correlation between a signer's head pose, in particular, the pitch angle of the head, and the presence of the wh-question expression, and used it to improve classifier performance, in a stacked SVM framework.

Bibliography

Athitsos, V., C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, et al. (2008). The American Sign Language Lexicon Video Dataset. *Proceedings of the IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis.*

Baker, C. (1980). Sentences in ASL. In C. Baker & R. Battison (Eds.), *Sign Language and the Deaf Community. Essays in honor of William C. Stokoe* (pp. 75-86). Silver Spring, MD: National Association of the Deaf.

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

Baker, C., & C. A. Padden (1978). Focusing on the nonmanual components of American Sign Language. In P. Siple (Ed.), *Understanding language through sign language research* (pp. 27-57). New York: Academic Press.

Baker-Shenk, C. (1983). *A Micro-analysis of the Nonmanual Components of Questions in American Sign Language.* Unpublished Doctoral Dissertation.

Baker-Shenk, C. (1985). The Facial Behavior of Deaf Signers: Evidence of a Complex Language. *American Annals of the Deaf, 130*(4), 297-304.

Bauer, B., & K.-F. Kraiss (2001). Towards an automatic sign language recognition system using subunits. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and Sign Language in Human-Computer Interaction. International Gesture Workshop* (Vol. 2298, pp. 64-75). New York, NY: Springer-Verlag.

Choi, J., W. J. Jeon, & S.-C. Lee (2008). Spatio-temporal pyramid matching for sports videos. In *MIT '08: Proceedings of the 1st ACM international conference on Multimedia information retrieval* (pp. 291-297): New York, NY: ACM.

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*(1), 38-59.

Coulter, G. R. (1979). *American Sign Language Typology.* Unpublished Doctoral Dissertation, University of California, San Diego.

Dreuw, P., C. Neidle, V. Athitsos, S. Sclaroff, & H. Ney (2008). Benchmark Databases for Video-Based Automatic Sign Language Recognition. *Proceedings of the The sixth international Conference on Language Resources and Evaluation (LREC).* Morocco. May 2008.

Kanaujia, A., Y. Huang, & D. Metaxas (2006). Tracking Facial Features Using Mixture of Point Distribution Models. *Proceedings ICVGIP 2006*.

Lazebnik, S., C. Schmid, & J. Ponce (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2169-2178.

Liddell, S. K. (1980). *American Sign Language Syntax*. The Hague: Mouton.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

MacLaughlin, D., C. Neidle, & D. Greenfield (2000). SignStream™ User's Guide, Version 2.0. Boston, MA: American Sign Language Linguistic Research Project No. 9, Boston University.

Neidle, C. (2002a). SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. Boston, MA: American Sign Language Linguistic Research Project Report No. 11, Boston University.

Neidle, C. (2002b). SignStream™: A Database Tool for Research on Visual-Gestural Language. *Journal of Sign Language and Linguistics, 4*(1/2), 203-214.

Neidle, C. (2003). *SignStream™ - Version 2.2 CD-ROM*.

Neidle, C. (2007). SignStream™ Annotation: Addendum to Conventions used for the American Sign Language Linguistic Research Project. Boston, MA: American Sign Language Linguistic Research Project Report No. 13, Boston University.

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

Neidle, C., S. Sclaroff, & V. Athitsos (2001). SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers, 33*(3), 311-320.

Shi, J., & C. Tomasi (1994). Good features to track. *CVPR*, 593-600.

Stokoe, W. C. (1960). Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf (Studies in Linguistics Occasional Papers, volume 8. No. Department of Anthropology and Linguistics, University of Buffalo). Buffalo, NY.

Tsai, C.-F., & C. Hung (2008). Automatically annotating images with key words: A review of image annotation systems. *Recent Patents on Computer Science, 1*, 55-68.

Vogler, C., & S. Goldenstein (2008). Facial movement analysis in ASL. *Universal Access in the Information Society, 6*(4), 363-374.

Vogler, C., & D. Metaxas (2004). Handshapes and movements: Multiple-channel ASL recognition. *Springer Lecture Notes in Artificial Intelligence (Proceedings of the Gesture Workshop '03, Genova, Italy), 2915*, 247-258.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*, 241-259.